Volume 43 Number 4 December 2019 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Editorial Boards

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Predictive Analytics on Big Data - an Overview

Gayathri Nagarajan and Dhinesh Babu L.D
School of Information Technology and Engineering, Vellore Institute Of Technology, Vellore, India
gayunagarajan1083@gmail.com
E-mail: lddhineshbabu@gmail.com

Overview paper

Keywords: predictive analytics, big data, machine learning

Received: November 5, 2018

Big data generated in different domains and industries are voluminous and the velocity at which they are generated is pretty high. While research works carried out continuously to handle big data is at one end, processing it to develop the business insights is a hot topic to work on the other end. Though there are lot of technologies and tools developed to handle big data and extract insights from them, there are lot of challenges and open issues that are yet to be addressed. This paper presents an overview on predictive analytics with big data. The overview throws light on the core predictive models, challenges of these models on big data, research gaps in several domain sectors and using different techniques. This paper categorizes the major technical challenges of predictive analytics on big data under six headings. The paper concludes with the identification of open issues and future directions to work on for researchers in this field.

Povzetek: Pregledni članek opisuje prediktivno analitiko na velikih podatkih.

1 Introduction

Research focus on predictive analytics for big data has gained significance because of its scope in various domains and industries. It has stepped into every field including health care, telecommunication, education, marketing, business, etc. Predictive analytics is a common diction that often means predicting the outcome of a particular event. The main idea behind prediction is to take certain input data, apply statistical techniques and predict the outcome of an event. The terminology 'predictive analytics' is synonymous with other terminologies like 'machine learning', 'data mining', 'business intelligence' and recently the other terminology which is in common use today 'data science'. Though they seem to be synonymous there is a narrow line that distinguishes their context of use.

The technique of business understanding, data understanding, data integration, data preparation, building a model to extract hidden insights, evaluating the model and finally deploying the model is called 'Data mining'. The model may be predictive or may not be. [1]. In some cases it may be descriptive whereas 'predictive analytics' in most cases mean to predict the value of certain output variable from input variables. 'Machine learning' is basically a technique whereas 'predictive analytics' is an application of machine learning. 'Machine learning' is used to discover hidden patterns in data by using some of their techniques like classification, association or clustering in training the machine. 'Machine learning' is one disciplinary of 'data mining' which is multidisciplinary that includes other disciplines like statistics, BI tools, etc. 'Data science' can be considered as an application of statistical methods to business problems. Predictive analytics is more narrowly focused than data science. Data science uses data programming whereas predictive analytics uses modeling. Predictive analytics in most of the cases is probabilistic in nature whereas data science involves exploration of data. Data scientists require both domain knowledge and the knowledge in technology. Business intelligence provides standard business reports, ad hoc reports on past data based on OLAP and looks at the static part of the data. Predictive analytics requires statistical analysis, forecasting, and causal analysis, text mining and related techniques to meet the need of forward looking business [1].

In predictive analytics, data is collected from different input sources. A model is developed based on statistics. The model is used to predict the outcome after proper validation. With the advent of big data, predictive analytics on big data has become a significant area of interest. Though there are lot of tools and techniques available to handle predictive analytics on big data, there are yet challenges open for the researchers to work upon. Our paper aims to present an overview on predictive analytics in big data to aid the researchers understand the contemporary works done in this area thereby providing them research directions for future work. We focused on including the research works carried in different industries and using different techniques so that the researchers can focus more on their specific area of interest after a complete understanding of the works done in different fields and using different techniques. The mo-

tivation behind this work is the fact that many papers in this field are more focused on a particular domain or technique but there is a lack of papers that presents a broader overview of predictive analytics in big data to help the budding researchers identify research problems. Hence we focussed on a comprehensive overview on predictive analytics.

This paper is organized into 7 sections. Core predictive models with their strengths, weaknesses along with few solutions are discussed in Section 2, the challenges of core predictive models on big data is discussed in Section 3, scope of predictive analytics on big data generated across different domain sectors along with few research gaps is discussed in Section 4 and the comprehensive challenges for predictive analytics on big data and the techniques used to overcome them is discussed in Section 5, the future directions for research are summarized in Section 6 and Section 7 winds up with conclusion.

2 Core predictive models

The major processes of predictive analytics include descriptive analysis on data that constitutes around 50% of the work, data preparation(like detecting outliers) that constitutes around 40% of the work, data modeling that constitutes around 4% of the work and evaluating the model that constitutes around 6% of the work [98]. Only a fraction of raw data is considered for building the model which is assessed and tested [7]. The phases involved in building predictive models is shown in figure 1. Initially predictive analytics was carried out using many mathematical statistical approaches. Later data mining, machine learning began its era in predictive analytics since they proved to be effective. This section discusses few core predictive models to make the reader understand the concept of predictive analytics. Different models are used for different types of predictive tasks such as estimating an unknown value, classification(supervised learning to predict the class label of the instance), clustering(unsupervised learning to group similar patterns together) etc. The section is branched into three subsections - the predictive models based on mathematical(statistical) approaches, the models based on data mining approaches and the models based on machine learning approaches respectively. Yet, there is a very narrow line of separation among the subsections and they overlap in certain predictive tasks. Figure 2 shows the classification of core predictive models.

2.1 Predictive models based on mathematics

Mathematical techniques especially statistics is used for predictive tasks. Despite, data mining algorithms and machine learning algorithms also use math as their base for predictive tasks. Major core predictive models based on mathematics include Extrapolation, Regression, Bayesian statistics that are described in detail.

2.1.1 Extrapolation

Extrapolation is a method of extending the value of a variable beyond the original observation range. For example, the details of the road condition are known to a driver until a certain point and he is expected to predict the road condition beyond that point. A tangent line is drawn by relating the input variable to the output variable. The line is extended further to predict the output for different values of input. The line determines whether the extrapolation is linear, quadratic or cubic etc.

Strength and weakness :

Extrapolation suits well for such tasks where the target variable is in close relationship with the predictor variables. The results of extrapolation are also accurate in certain experiments where the relationships among the variables are simple [101].

The major problem with extrapolation is the interpretation of results. There are many studies where the study population differs widely from the target population.[100] is an example of such problem where the extrapolation of the experimental results on sheep cannot be justified for other target population. In such studies, the claims of the study results cannot be applied or justified to the target population [99]. Few solutions proposed to solve the interpretation problem of extrapolation is simple induction, randomized trails and expertise, Mechanistic reason etc. Population modeling is also proposed to solve the extrapolation problem [102]. But these solutions can help only to a certain extent. Secondly, it is hard to model the past with extrapolation. Sometimes several extrapolation methods are combined to model. Moreover, extrapolation cannot be used to model the tasks with non linear patterns [103].

2.1.2 Regression

Regression models are used for supervised learning problems in predictive analytics. It works by establishing a mathematical relation of the input variables with the output variable. There are different types of regression models like linear regression model, multi variate regression model, logistic regression model, time series regression model, survival analysis, etc. depending on the nature of the relationship discovered among the variables. Though the term is synonymous with extrapolation, there is a difference. Regression explains the variations in the dependent attribute with respect to the variations in the predictor attributes. Also, regression doesn‘t use the value of the input variables outside the range to establish the relationship with the dependent variable as in the case of extrapolation. There are different variants of regression depending on the nature of the variables as shown in table 1.

Strength and weakness:

Linear regression can suit well on tasks where the variables exhibit linear relationship [104]. For predictive tasks where associated probability is also important apart from predicting the value of a variable such as in [110], logistic regression is preferred. For predictive tasks where a

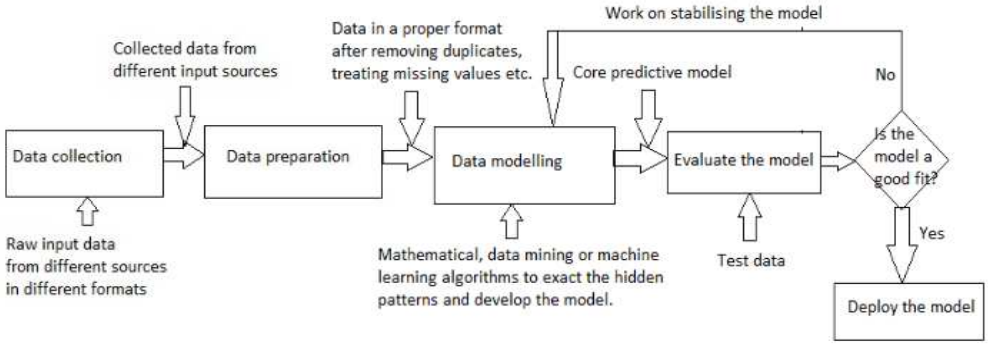


Figure 1: Phases involved in building a core predictive model

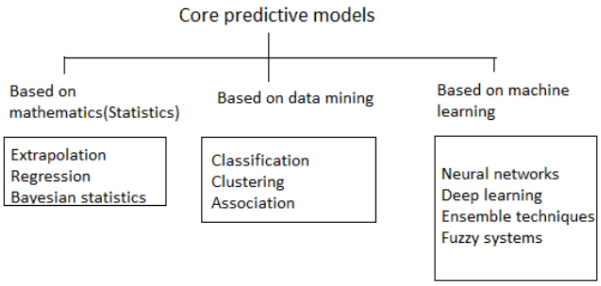


Figure 2: Classification of core predictive models

non parametric and non linear methodology has to be used, MARS regression can be considered as it does not assume any functional relationship between the dependent and independent variables [114],[117]. The biggest strength of MARS is that it is simple to understand and easy to interpret. Despite, it does not require any data preparation [116]. Moreover it is suitable for problems with higher input dimensions due to its 'divide and conquer' strategy of working principle [117]. To model complex processes, in which no theoretical models exist, LOESS regression is preferred as it does not require a function to fit a model to all data in the sample [121]. The major strength of LOESS regression is its flexibility. The flexibility is controlled by a smoothing parameter [119]. It is usually set between 0.25 and 0.5 [120]. For predictive tasks, where the number of predictors is more, regularization methods such as ridge regression is preferred as it avoids overfitting [122]. It also handles multicollinearity effectively [126].

The major weakness of linear regression is that it considers the mean of dependent variable and hence is sensitive to outliers. It is also more suited only to applications in which the predictor variables are independent [104]. For example, [105] states the reasons for moving to machine learning predictive models from simple regression models for predictive tasks in medicine. Though the regression models are simple and robust, they are limited to a small number of predictors that operate within a range. Outlier detection algorithms are proposed for linear time series regression models [106],[107]. Another problem with linear regression models is that it does not suit for predictive tasks when the predictor variables are collinear. [109] is an example that shows the adverse effect in interpretation of results when regression is performed without considering the multicollinearity problem. Techniques such as principal component analysis or stepwise regression can help to remove highly correlated predictors from the model [108]. The major weakness with logistic regression is that it is affected by omitted variables. Solutions such as replacing the latent continuous variable with an observed continuous one is proposed. Moreover,the odds ratio obtained from logistic regression cannot be interpreted easily as the heterogeneity of the model is not accounted [112]. In clinical predictive tasks, the odds ratios are estimated as risk ratios which is actually an overestimation. Alternatives such as Mantel–Haenszel risk ratio method, log–binomial regression, Poisson regression are used to give correct risk ratios [113]. Interpretation of results can be achieved to a certain extent by observing the probability changes [111]. Logistic regression is also not found to perform well with class imbalance problems and in such cases algorithmic approaches such as random forests is used [112]. The major weakness of MARS regression is overfitting and methods such as generalized cross validation is used [115],[118]. Pruning technique is also used to avoid overfitting problem to some extent by reducing the number of its basic functions thereby limiting the complexity of the model [117]. LOESS regression is less sensitive to outliers yet they are also overcome by extreme outliers. Another disadvantage with LOESS regression is that it requires densely sampled data to produce good results [121]. The major problem with ridge regression is the parameter settings. The hyperparameter has to be set [124]. Few methods are proposed in [123], [128] for parameter setting. Ridge regression also suffers from interpretability [124]. This happens because the unimportant predictors still exists in the model with the coefficients close to zero but not exactly zero [125]. LASSO regression is preferred in such predictive tasks to avoid the inter-

S.no	Type of regression	Explanation
1	Linear regression	The linear regression is given as $$Y = aX + b \quad (1)$$ where Y represents the predictor attribute, X represents the independent attribute, a is the slope and b, the intercept. The best fit line is obtained by minimizing the square of variations of each observed and the actual values with the line.
2	Logistic regression	Logistic regression is used for problems that are binary in nature and hence is mainly used for classification. This method aids to determine the likelihood of the occurence of a happening. It is denoted by $$logit(a) = ln(\frac{a}{1-a}) \quad (2)$$ where a is the likelihood of the occurence of the event.
3	MARS	Multivariate adaptive regression splines represented as MARS uses stepwise regression for model building. It is a non parametric method. The non linearities among the variables are identified and modelled with hinge functions. These functions create a hinge in the best fit line automatically according to the non linear relationship among the variables. The MARS equation is given as $$D = \beta_0 + \sum_{n=1}^{N} \beta_n h_n(I) \quad (3)$$ where D represents dependent attribute, I represents independent attribute, β_0 represents intercept variable, β_n represents slope of the hinge function $h_n(I)$.
4	LOESS regression	LOESS regression is a non parametric regression model. This method helps to fit regression line on subset of data rather than the data as a whole. It incorporates the concepts of regression model along with the nearest neighbor concepts.
5	Ridge regression	Ridge regression is another method commonly applied when the dataset experiences multicollinearity. They reduce the standard errors. A shrinkage parameter λ is added to the least squares term to minimize the variance.Ridge regression estimator is given as $$\beta_{ridge} = (I^T I + \lambda I_p)^{-1} I^T D \quad (4)$$ where I represents independent attribute matrix, D represents predicted attribute matrix, Ip represents identity matrix and λ represents shrinkage parameter.

Table 1: Different types of regression

pretability problem of ridge regression as it sets the coefficients of unimportant parameters to zero [127].

2.1.3 Bayesian statistics

Bayesian statistics predicts the likelihood of events occurring in the future. It works in the same way like normal probability but uses the input from experimentation and research to adjust the beliefs. For example, the probability of a six occurring in a die thrown six times is 1/6. But Bayesian statistics starts with the initial value of 16.6% and then adjusts the belief based on the experimentation. If the die is showing 6 more than the expected number of times during experimentation, the value of this belief is increased accordingly. Hence the likelihood of 6 turning in a die thrown 6 times will increase or decrease depending on the outcomes in the experimentation.

Strength and weakness:

Bayesian approaches yield accurate results in many predictive tasks as they consider both experimental data and theoretical predictions [129]. Bayesian approaches are best suited for tasks where there are uncertainties in models and parameters. They also find their use in predictive tasks where probabilities questions have to be answered such as

in stock market analysis [130].

The major weakness of bayesian approaches lies in determining the prior distributions. Bayesian approaches are also computationally expensive [130]. Use of frequencies instead of probabilities can help in improving bayesian reasoning [131].

2.2 Predictive models based on data mining

The process of extracting hidden patterns from the given input is called data mining. It is basically mining knowledge from the data. Three major approaches for data mining include Classification, Clustering and Association. Machine learning algorithms are widely used to execute the task.

2.2.1 Classification

The method of determining the class to which a particular data given as input belongs to is called classification. It is a supervised machine learning technique with labelled input data. Classification can be used in predictive analytics. There are lots of algorithms under classification technique but few basic algorithms are described in detail in this subsection.

1. Naive Bayes: Naive Bayes is a statistical modeling approach with two assumptions —all the attributes are equally important, all the attributes are statistically independent. It is a probabilistic approach which works on the following Bayes theorem.

$$P[\mathrm{M}/\mathrm{N}] = \frac{P[\mathrm{N}/\mathrm{M}]P[\mathrm{M}]}{P[\mathrm{N}]} \tag{5}$$

Strength and weakness:

The main strength of naive bayes is its simplicity. Inspite of the fact that its accuracy is less, it is found to perform better due to its simplicity in tasks such as document classification where merely classification is important. Naive bayes is computationally efficient since the contribution of each variable is equal and independent [134]. Moreeover only few parameters need to be calculated in naive bayes due to its conditional independence assumption. Hence it suits well for tasks where the training data is less [135].

The major weakness of naive bayes approach is its conditional independence assumption that often does not hold for real world applications [132]. This weakness is overcome to a certain extent by weighting attributes. Naive bayes with attribute weighting is found to perform better than random forest and logistic regression in [136],[137]. Accuracy and speed of classification is also less when compared with other classification approaches. Effective negation and feature selection techniques such as mutual information in combination with naive bayes is found to improve the accuracy and speed to a certain extent[133]. An another problem with naive bayes is that though they are good classifiers, they are not good estimators as discussed earlier. Hence it does not perform well in the tasks where probability value is important [138] and certain improvements to naive bayes is proposed to improve the probability estimation [139]. Moreover when the test data differs widely from the training data naive bayes fails to perform well unless smoothing techniques such as laplace estimation is used [138].

2. Decision trees: Decision tree is constructing a tree based structure for classification. Each node involves testing a particular attribute and the leaves are assigned classification labels based on the values at each node. Decision trees use divide and conquer approach and the attributes can also be selected with heuristic knowledge for the nodes of decision trees though few measurements like entropy and information gain are used in selecting the attributes. Decision trees can be converted to rules also. Many variations of decision trees have evolved and one of them is random forest which is commonly used bagging method in recent research problems. The leaf nodes of the decision trees are called decision nodes. Entropy is the amount of randomness in the data. Information gain is the information obtained that helps for accurate prediction. Entropy is given by

$$E(X) = -\sum_{i=1}^{n}(P_i log P_i) \tag{6}$$

where $\mathrm{P_i}$ is the probability of occurence of value i when there are n different values.

Information gain is a purity measure given by

$$IG(X, a) = E(X) - E(X|A) \tag{7}$$

The value represents the information gained by splitting the tree with that particular attribute. The attribute with less entropy or more information gain at every stage is considered the best split and the process is repeated until the tree converges. There are many decision tree algorithms but few variants are shown in table 2.

Strength and weakness:

The major advantage of decision tree over other classifiers is its interpretability. The tree like structure helps users to extract the knowledge easily [140]. Indeed, decision trees does not require the data to be normally distributed. The data can be continuous, discrete or a combination of both. Hence there is no need for much data preparation in decision tree model [142]. Moreover decision trees require only very few training iterations [143]. The random forest, an ensemble technique of decision tree is found to yield more accurate results. An another advantage of random forest is that it is non parametric in nature and helps in determining variable importance [141].

S.no	Type of decision tree	Description
1	ID3	Iterative dichotomiser is the basic non incremental algorithm used for construction of decision tree. It uses information gain measure to select the best split at each node. But the major disadvantage of ID3 is that it may overfit the training data and may not result in optimal solution.
2	C4.5	C4.5 is an improved version of ID3 algorithm. It solves the overfitting problem of ID3 by pruning the tree. C4.5 handles both continuous and discrete attributes and is capable of handling missing values too.
3	C5.0	C5.0 is an improved version of C4.5 in terms of performance and memory efficiency. It supports boosting technique to improve accuracy. C5.0 constructs smaller decision trees by removing unhelpful attributes without compromising on accuracy.
4	Decision stumps	It is a single level decision tree and finds its use along with machine learning techniques like bagging and boosting as weak learners.
5	CHAID	Chi square automatic interaction detector is used to find the relationship between categorical dependent variable and categorical independent variables. It uses chi square test of independence to test the independency between two categorical variables. CHAID can be extended for continuous variables too.

Table 2: Different types of decision trees

The major problem with decision trees is overfitting or underfitting [144]. Techniques such as pruning [146],[147] or feature selection methods are required to avoid overfitting problem and also to reduce the computational complexity of the model [147]. Decision trees does not suit well for imbalanced datasets though ensemble techniques can help [145]. Moreover, though random forest yields more accurate results, they are black box classifiers as the split rules are unknown [141].

3. KNN: Another classification technique that is of wide use is KNN yet with its own challenges. This technique works on the idea that the input data to be classified depends on the class of its neighbors. The value of 'K' determines the effectiveness of the algorithm. 'K' represents the number of neighbors to be considered. The input data is assigned to the class to which most of its neighbors belong to. A distance metric from the input data to the 'K ' neighbors is calculated. Euclidean distance is usually deployed to calculate the distance. Other distance metrics like mahanalobis and manhattan distance measures can also be used instead of euclidean. The accuracy of the algorithm lies in the choice of K. Lower value of K might result in overfitting and higher value for K might result in a more generalized model difficult to predict.

   Strength and weakness:

   The biggest advantage of KNN is its simplicity [150]. It does not require any prior knowledge about the distribution of data [149]. This non parametric nature of KNN makes it effective for real world problems [151].

   The major issue with KNN is the choice of parameter k and distance metric [150],[152],[153] and few works are proposed to determine the value of k [157],[158] etc. Computational complexity is another issue with KNN. Techniques such as clustering are used along with KNN [148] to reduce the computational complexity. KNN is also affected by irrelevant features [150] and is sensitive to outliers [152],[153]. The outlier problem can be avoided to a certain extent by choosing a reasonable value for k rather than a small k [154]. Few methods or improvements such as local mean based KNN [155] and distance weight KNN [156] are proposed to overcome the negative effect of outliers in KNN.

4. Support vector machines: This classification technique yields better accuracy in classification problems. SVM works on the basis of hyperplane. The idea behind this technique is to find the best hyperplane that can classify the two classes more accurately. This technique is best suited for both linear and non-linear separation problems. Non-linear problems can be handled using kernel functions that does data transformations to find the best hyperplane classifying the data more accurately. This algorithm works under the concept of margin. The distance between the hyperplane and the closest object in each class is calculated. The hyperplane with maximum margin with the class objects is the best classifier since it can predict the class more accurately. Each input data is assigned a point in n-dimensional space.

   Strength and weakness:

   The major strength of SVM is its robustness. It models non linear relationships very effectively and hence is proved to yield better results in terms of accuracy especially in non-linear problems [161][165]. SVM

is also known for its generalization capability and the generalization error is less in SVM [163] [164]. This advantage of SVM helps it to model complex problems even when the training data is less [166]. Moreover, there is no need for feature extraction process in SVM as the kernel function can be directly applied on the data [164]. SVM also avoids overfitting [165], [166].

The major weakness of SVM lies in its parameter settings. Proper setting of kernel parameters determine the accuracy of SVM. Certain optimization techniques such as PSO [159] and GA [160] are used to optimize the parameters of SVM. Methods such as double linear and grid search are also used to determine the parameter values of SVM [162].

2.2.2 Clustering

The technique of identifying similar patterns in the input data and grouping the input data with similar patterns together is called clustering. Clustering helps in predictive analytics. An example of clustering algorithm includes segmenting customers based on their buying behavior pattern thereby helping to predict the insights in the business and improve the sales accordingly. Clustering the scan images in health care helps to predict whether the person is affected by a specific disease or not. Though there are many clustering algorithms, the three basic algorithms include k-means, hierarchical and density clustering and a summary of the same is provided in the following subsection. A review of clustering with its scope, motivation, techniques and applications are explained in [2].

1. K-means clustering: K-means clustering chooses K random centroids and measures the distance of each input data point with the centroids. The most commonly used distance measurement metric is Euclidean distance. The input data points within the specific distance from the centroid are grouped together as a cluster and hence arrived at few clusters. The average of the distance of all the points from the centroid inside a cluster is calculated and the centroid is recalculated accordingly. The input data points belonging to the cluster changes again. This process continues until the centroids are fixed. Another variation of partition clustering is K-mediods where the centroid itself is an input data point. K-median and K-mode algorithms are also partition based clustering algorithms that uses median and mode instead of mean. There are several metrics to measure the performance of clustering. One among them is the distance metric. Single linkage is the nearest neighbor clustering where the minimum distance between the data points of two different clusters is calculated. Complete linkage is the farthest clustering where the maximum distance between the data points of two different clusters is calculated. Average linkage is also used in some scenarios.

   Strength and weakness:

   The major strength of k means clustering is it‘s simplicity and speed [168]. It can also work on datasets containing a variety of probability distribution [175].

   The major drawback with k means clustering is it‘s sensitivity to the initialization of cluster centers [167]. Hence determining the initial cluster centers is a major challenge though many methods based on statistics, information theory and goodness of fit are proposed [168]. Determining the number of centers is also a challenge and is addressed in few works [173]. An another drawback with k means clustering is it‘s computational complexity. As the distance of each data point has to be calculated with each cluster center for every iteration, the computational time is high. Solutions such as data structure that stores information at each iteration to avoid repeated computations [169] and Graphical processing units (GPUs) that parallelize the algorithm are proposed to reduce the computational complexity of k means clustering [172]. Moreover k means clustering is also sensitive to outliers and can end up in local optima. Few alternatives include fuzzy c means clustering and other soft clustering techniques that are proved to work well with noisy data [170]. k means clustering is also combined with optimization techniques such as PSO and ACO to avoid local optima and to arrive at better cluster partitions [171]. Few works are carried out to identify the better cluster partitions with minimum intracluster distances and maximum intercluster distances. Optimization function is derived that minimizes the intracluster distances and maximizes the intercluster distances. This function is optimized using optimization algorithms such as GA, PSO, WOA, ACO etc in few clustering works. Few other works include [269] that uses a set of objective functions and updates the algorithm accordingly to improve the intracluster compactness and intercluster separation, [270] that uses bisected clustering algorithm to measure the intracluster and intercluster similarity metrics etc. [174] proposes a method to overcome the drawback of noisy features in k means clustering.

2. Hierarchical based clustering: Hierarchical clustering works either in a divisive way (top-down) or agglomerative way (bottom-up). In the divisive clustering, large cluster is broken down into smaller pieces. In the agglomerative clustering, each observation is started as its own cluster and pair of clusters is merged together as they move up in the hierarchy. A dendogram is a pictorial representation for hierarchical based clustering. The height of the dendogram represents the distance between the clusters. Agglometric and divisive clustering algorithms are called AGNES and DIANA respectively. In DIANA clustering technique, all the input data points are considered as a single cluster and every iteration divides the cluster based

on heterogeneity. More heterogeneous data points breaks down into another cluster. In AGNES clustering technique, each input point is considered as a single cluster and homogeneous points are clustered together as a single cluster at each iteration.

Strength and weakness:

The major strength of hierarchical clustering includes it‘s scalability and capacity to work with datasets of arbitrary shapes [177]. It also determines the hierarchical relationships well. Moreover the number of clusters need not be specified in advance [178].

The major drawback with hierarchical clustering is its computational complexity [176],[177],[178] and few other methods are proposed to improve the efficiency of the same [176]. Parallel techniques are also used to improve the computational efficiency of hierarchical clustering [179].

3. Density based clustering: Density based clustering works by defining a cluster as the maximal set of density connected points. Three types of points are chosen core, border and outlier. Core is the part of the cluster that contains dense neighborhood. Border doesn‘t have many points but can be reached by the cluster. Outlier can‘t be reached by the cluster. Density based clustering picks up a random point and checks if it is the core point. If not, the point is marked as an outlier. Else, all the directly reachable nodes from the specific point are assigned to the cluster. It keeps finding the neighbors until it is unable to. There are certain kinds of problems where density based clustering provide accurate results than k-means clustering. Outlier detection is accurate in density based clustering.

Strength and weakness:

The major strength of density based clustering is that it can discover clusters of arbitrary shapes [184],[177]. It is also robust to outliers [184]. There are several density based algorithms such as DBSCAN, OPTICS, Mean-Shift etc [177].

The major drawback with density based clustering is the setting of parameters. Parameters such as neighbourhood size, radius etc. have to be set in density based clustering [182], [177]. Few algorithms are proposed to determine the parameters in density based clustering [183]. Moreover the density of the starting objects affect the behavior of the algorithm. The algorithm also finds its difficulty in identifying the adjacent clusters of different densities [182] ,[177]. Techniques such as space stratification is proposed to solve this problem [182]. An another drawback with density based clustering is its efficiency. Parallelization of the algorithm reduces the computational complexity to a certain extent. Techniques such as GPUs [180], mapreduce [181] are used to improve the scalability and efficiency of the algorithm.

2.2.3 Association

The method of identifying the relationship amid the items and deriving the rules based on the relationship is called association. Though association mining is not of much use in prediction, there are few scenarios where association rule is used. The rule has antecedent and consequent. Association rule mining is used mainly in business and marketing [185]. There are different algorithms used in association rule mining. Few include Apriori, Predictive Apriori, Tertius etc [186]. Optimization techniques such as PSO are also used with association rule mining to improve its efficiency [187]. [188] presents a survey on association rule mining algorithms.

2.3 Predictive models based on machine learning

Machine learning approaches are used for predictive tasks. It is the process of training the machine with a training input set, building a model and the evaluating it with the test data. The machine learns continuously from the errors until the model gets stabilized. Supervised learning works with labeled input data whereas unsupervised learning works with unlabeled input data. Machine learning uses soft computing techniques like neural networks for training.

2.3.1 Neural networks

Neural network is a commonly used soft computing technique for predictive analytics. Neural networks are used to classify complex patterns that are difficult to classify using SVMs or other techniques. There are different types of neural networks that can be trained using supervised, unsupervised and reinforcement learning. There are also different learning algorithms for training neural networks.

Neural networks machine learning algorithm can be used to train a network with a group of training data and then test it with a group of test data thereby measuring the accuracy of prediction. Learning continues until the network becomes stable and able to classify the data accurately. Cross validation is one among the widely used technique for evaluating the model. Backpropagation algorithm is the most commonly used training algorithm in neural networks. Weights are assigned at each layer input, hidden and the output. Weightage is given to each attribute based on the impact of it in predicting the output variable. Different types of functions like sigmoidal function and sign function are used to compute the output variable. These functions are called threshold functions and the output variable is predicted based upon these functions. Threshold functions are also called activation function or transfer function. The choice on number of input nodes, hidden layers, weightage, threshold functions, algorithm for learning are all based on the application and data for which predictive analytics has to be applied. Now, deep learning techniques are used to improve accuracy.

Strength and weakness:

The major strength of Artificial Neural Networks(ANN) lies in it‘s ability to work with large amounts of data and yield good results. They have good generalization and learning ability and are universal approximators [191]. ANN has good fault tolerant, self learning, adaptation and organization ability [192]. An another advantage with ANN is that they are good for high dimensional datasets as each variable do not have major impact on the class variable but as a group they are good at classification. Moreover a complex ANN relives user from determining the interactional and functional forms in prior and is able to smooth any polynomial function [193]. There are different types of neural networks such as as feedforward network, radial basis function network(RBFN), auto encoder, Boltzmann machine, extreme learning machines, deep belief network, deep convolutional network etc [189] each with its own strengths and weaknesses. For example, RBFN are easy to design, have good generalization ability and are tolerant to noise. These networks find their use in designing flexible structures and systems [190].

The major weakness with ANN is that they can‘t be applied blindly to all kinds of data. Hence they are used by combining with other models as hybrid prediction models in most of the prediction problems. For example, in time series problems, both linear and non linear relationships exist and ANN is combined with ARIMA modelling in such problems [190]. An another disadvantage lies in the fact that there are no proper rules to determine the number of hidden nodes in neural networks. Moreover they can also easily end up in local optima and are tend to overfit [193]. Optimization algorithms such as GA [194], Gravitational search algorithm with PSO [196] are used to avoid the local optima problem in ANN. Algorithms such as Fruitfly algorithm also find their use in determining the parameters for ANN [195]. Overfitting problems is addressed by techniques such as dropout mechanisms [197], bayesian regularization [198] etc.

2.3.2 Deep learning

Deep learning is the most commonly used technique in use today for classification, clustering, prediction and other purposes. While learning in machine learning proceeds in a broader way, deep learning works in a narrow way. It works by breaking down the complex patterns into simple smaller patterns. Learning happens in parallel in the smaller patterns and finally the sub solutions are combined together to generate the final solution. This improves the accuracy of the network. Deep nets also help in avoiding the vanishing gradient problem. Most of the deep learning problems use Rectified Linear units function(ReLUs) instead of sigmoidal and tanh activation functions that causes vanishing gradient problem. The use of ReLUs help overcome the vanishing gradient problem by avoiding zero value for the derivative and maintaining a constant value instead [271]. Moreover the use of deep learning networks such as Long Short Term Memory Networks(LSTM) avoid vanishing gradient problem by maintaining a constant value for the recursive derivative using cell states [272]. Deep nets use GPUs that help them get trained faster. When the input pattern is quite complex, normal neural networks might not be effective because the number of layers required might grow exponentially. Deep nets work effectively in solving such complex pattern by breaking them down into simpler patterns and reconstructing the complex solution from the simpler solutions. GPUs are known to be a better alternative to CPUs for big data analytics owing to it‘s lower energy consumption and parallel processing. Hence GPUs are found to be scalable in deep learning as the training and learning of the deep nets are made faster with parallel processing of many computational operations such as matrix multiplications [273]. There are different kinds of deep nets used for different purposes [5]. Table 3 shows the different types of deep nets and their usage.

Strength and weakness:

An overview of deep learning in neural networks has been discussed in [199]. The major strength of deep learning is it‘s ability to model non linear relationships well. Deep learning also suits well for massive data and has better representation ability than normal networks [200]. Moreover deep learning does automatic feature extraction and selection [201].

The major weakness of deep learning is that it is a black box technique. There is no theoretical understanding behind the model. Certain techniques such as information plane visualization are proposed to understand DNN by using the mutual information exchanged between layers in DNN [202]. Moreover deep learning works well only with massive data and their architectures are more specialized to a particular domain. They also consume high power [203].

2.3.3 Fuzzy rule based prediction

Fuzzy logic is a concept of soft computing technique more suited for prediction problems with uncertainty and imprecision. Fuzzy sets have membership functions associated with each input data set. The membership value of a particular input data represents the level of belonging of the particular input data to the particular set. Rules are derived and learning is based on the rules. Finally the rule based approach is used to classify or predict the output variable. Fuzzy systems are widely used for prediction purposes. Fuzzy systems can be used as stand-alone or can also be combined with other machine learning algorithms for predictive tasks. The simple fuzzy based classifier is If-THEN classifier and it can be made more meaningful with the use of linguistic labels [61],[204]. Fuzzy systems are also combined with neural networks as neuro-fuzzy classifier [261],[209] and is used for prediction purposes. Fuzzy systems are also combined with KNN for prediction purposes. Fuzzy c means clustering is found to perform well than hard clustering especially in applications such as bioinformatics where genes are associated with many clus-

S.no	Type of deep net	Description	Usage
1	Restricted Boltzmann machine	Two layered network with visible and hidden layer. Layers not connected among themselves. In the forward pass, RBM takes the input and encodes as numbers. The backward pass does the reverse. Data need not be labelled.	Recognize inherent patterns. Works well with real time data like photos, videos, voice etc. Used for clustering.
2	Deep belief nets	Stack of RBMs arranged together. Output of hidden layer of the previous networks like RBN is given as input to visible layer of next RBN.	Used for recognizing complex patterns. Used more commonly in facial recognition.
3	Convolution nets	Made up of three layers, convolution, RELU and pooling each having its own function.	Used to identify the internal patterns within an image.
4	Recurrent networks	A network with built in feedback loop. Uses techniques like Gating to overcome vanishing gradient problem.	Used when the patterns in the input data changes over time. For image captioning, document classification, classify videos frame by frame, natural language processing etc. LSTM is a recurrent network architecture that is used for deep learning. The application of LSTM includes time series data predictions, classification, processing, hand writing recognition, speech recognition etc. It is known for reducing the exploding and vanishing gradient problems.
5	Autoencoders	Encode the input data and reconstruct it to back. Works with unlabeled data.	Finds its use in dimensionality reduction. Used for text analytics.
6	Recursive Neural Tensor nodes	Works with the concept of roots and leaves. Data moves in the network in a recursive way.	Used to discover hierarchical structure of a set of data. Used in sentimental analysis. Used in natural language processing.
7	Generative adversial networks	A network that can produce or generate new data with the same statistics as the training data[274].	Used in fashion designing, improving satellite images, etc.

Table 3: Different types of deepnets and their usage

ters [170], [270].

Strength and weakness:

The major strength of fuzzy systems is it‘s interpretability. The fuzzy models are easy to interpret if designed carefully [205] especially with the use of linguistic labels [206]. Fuzzy rules also help to model the real world processes easily [207]. Fuzzy systems are known well for handling uncertainty [208].

The major weakness of fuzzy systems include it‘s poor generalization capability as it is rule based. Fuzzy systems are not robust as any change should be incorporated into the rule base. To overcome this disadvantage, fuzzy systems are often combined with ANN and hybrid systems are developed for prediction [208]. An another disadvantage of using fuzzy systems is that the knowledge about the problem should be known in advance. The use of hybrid systems can help overcome this disadvantage as the knowledge is extracted from neural networks in such systems [209]. Approaches such as genetic programming is also used to generate rules for fuzzy systems [210].

2.3.4 Ensemble algorithms

Ensemble methods are combination of more than one technique to achieve more accuracy in prediction than achieved by an individual model. Few ensemble techniques are shown in table 4. Each ensemble technique has its own strength and weakness. For example, bagging is stable against noise but needs comparable classifiers whereas boosting is unstable against noise but its classification performance is better than bagging [211]. Also, bagging is found to perform better than boosting for class imbalance problems especially in noisy environment [212]. Stacking has it‘s own weakness with respect to computational time. It is computationally expensive [213]. Another problem with stacking lies in the selection of base level classifiers as techniques such as exhaustive search consumes more time when search space is large. Yet, unlike bagging and boosting that uses the same algorithm, stacking uses a different algorithm and hence heterogeneous in nature [215]. The choice of the ensemble technique depends largely on the problem at hand. Bagging is good to deal with problems

S.no	Ensemble	Description
1	Bagging	Bagging or bootstrap aggregation is the method of decreasing the variance without any change in the bias. It is mainly used for regression and classification techniques. Each model is built separately and the net output is derived by bringing together the results from the individual models by joining, aggregation and other methods.
2	Boosting	Boosting is a parallel ensemble method to reduce bias without any changes in variance. Boosting converts weak learning algorithms to strong learning algorithms using certain methods like weighed average. There are many variations of boosting algorithms like adaboost, gradient boosting etc. The misclassified instances are assigned more weight in the successive iterations.
3	Stacking	Stacking is the technique in which the output of the previous level classifier is used as training data for the next classifier to approximate the target function. It minimizes variance and methods like logistic regression is used to combine the individual models.

Table 4: Ensemble techniques

where a single model is likely to overfit whereas boosting is good for problems where a model yields poor performance. Moreover bagging can be done in parallel as each model is independent whereas every model in boosting depends on the previous model [214]. There are different kinds of boosting techniques, the major include adaboost and gradient boost. Adaboosting improves performance by assigning high weight to the wrongly classified data points whereas gradient boosting improves performance by using gradients in the loss function [216]. Indeed, gradient boosting converges in a limit whereas adaboost is computationally efficient than gradient boosting [217].

3 Challenges of core predictive models on big data

'Big data' represents data sets that are in petabytes, zettabytes and Exabyte. The sources of big data include satellites that generate enormous information every second from space, mobile communications generating voluminous data, social media like Facebook, Twitter with blogs, posts etc. Traditional relational databases, data warehouses and many visualization tools and analytical tools are developed for structured data. Because of the heterogeneous nature of big data and enormous amount of data generated including real time data, there is a need to enrich traditional analytical methods to support the analytical functionalities for big data. Alternatively, new tools and techniques are developed to work on big data in combination with the traditional analytical techniques. Big data development includes the development in all the areas of handling big data including data storage, pre-processing, data visualization, data analytics, online transaction processing, online analytical processing, online real time processing, use of business intelligence tools for predicting insights from big data etc [3]. The main characteristics of big data include volume, velocity, variety, veracity and value.[4]. A single machine cannot store big data because of its volume. The basic concept behind big data storage is to have many nodes (computers) and store the chunks of big data in them. The nodes are arranged in racks and communicate with each other and the centralized node that controls them. Clouds are also used for big data storage but it has its own challenge of privacy and security. Getting into the depth of storage technology is outside the purview of this article and hence we are leaving the discussion about big data storage at this stage. As our paper mainly aims to discuss the overview of predictive analytics with big data, this section addresses the challenges encountered by the core predictive models discussed in previous section on big data.

Extrapolation

Extrapolation will be precise only when the knowledge about the underlying covariate information [220] and the actual system is clear [219],[221] which is difficult to determine in big datasets. With big data such as spatial data, existing extrapolation approaches fail due to it's time and space constraints. Hence new technological innovative approaches are required to model such big datasets and understand them [219]. Extrapolation with kernel methods like gaussian are proved to be good due to their flexibility in choosing the kernel function. Yet, when it comes to development of gaussian models for multidimensional big data, it suffers from computational constraints. Techniques such as recovering out of class kernels are used to overcome the computational constraint to a certain extent [218]. An another problem with the machine learning models including deep learning is that they merely fit the data and may perform well for training dataset and even testing dataset but fails in extrapolation [221],[222]. This happens as they do not have proper structural explanations for the correlations they identify [222]. [220],[221] recommends the construction of hybrid systems comprising the science based model or physical models along with the predictive models to improve the accuracy of extrapolation. But the problem in developing hybrid systems lies in the fact that it requires domain knowledge.

Regression

Regression is easy and can be understood well when the data is small and can be loaded into memory com-

pletely. But big datasets can not be loaded into memory completely. Few parallel techniques and solutions are proposed for regression yet they end up in local optima or in accessing the data again and again for updates. The other problem in using parallel techniques is the computational resources incurred [223]. The area in the improvement of computational resources is still lacking when compared with the amount of big data generated [224]. Regression approaches such as kringing is computationally complex especially with big data. Sampling techniques such as leveraging [224] and subdata selection[226] are proposed to reduce the computational complexity. But as discussed in the earlier sections, the inferences on the samples cannot be justified completely for the whole population as such. Regression is also performed locally by dividing the big dataset into few smaller datasets and then combining the submodels to construct the final model [225]. The challenges with these solutions lies in the choice of appropriate method for division, aggregation etc.

Decision trees

Big data streams are more prone to noise and decision trees are more sensitive to noisy data [227]. The time taken to build the decision tree is computationally expensive with big data [228]. Preprocessing and sampling the big data in full batches before the construction of decision tree adds to the computational cost [227]. External storage is required to construct decision tree for big data as the complete dataset cannot be loaded into memory. Hence tradition decision tree design does not suit for the big data. Solutions such as incrementally optimized decision tree algorithm [227] is proposed where decision tree is built incrementally. Parallel techniques [229] are proposed in big data platforms such as spark where the decision tree algorithm is executed in parallel. Decision tree algorithm is also converted into mapreduce procedures in [228] to reduce the computational time. The computational time of gradient boosted trees is decreased in [230] by eliminating few instances in calculation of information gain and bundling certain features together. Yet, these solutions come at the cost of choosing the right technique to break the algorithm for parallel execution, bundling the features etc.

K Nearest Neighbor

The major problem of KNN with big data is it‘s computational time as the distance has to be calculated among each instances [231]. This in turn incurs memory requirement for storage [232]. k means clustering is used to cluster the big dataset and KNN algorithm is used on each subset to reduce the computational time [231]. But this solution comes with the general limitations of k means clustering. Memory requirement is handled to a certain extent by big data platforms such as spark so that in time memory computation is used effectively [232]. Map reduce approaches [233] are also used to reduce the computational time. Parallelization of KNN algorithm is also proposed [234]. Yet, all these big data platform solutions come with their own concerns on the nature of partitioning as the accuracy can not be compromised for efficiency [235].

Naive Bayes

Naive Bayes requires the probability to be calculated for all the attributes. With big datasets, the number of attributes is more and hence the time complexity to calculate the probability for all the attributes is high [236]. Another problem with naive bayes is the underflow and overfitting problems [237]. The underflow problem is usually handled by calculating the sum of log of probabilities rather than multiplying the probabilities whereas overfitting problem is handled using techniques like laplace estimate, M-estimate etc. But with high dimensional big datasets like genomic datasets, these solutions are not efficient [237]. Naive bayes deals only with discrete data. Hence discretization methods are used before applying naive bayes algorithm. In case of big data, existing traditional discretization methods are not efficient and may lead to loss of information [238]. Parallel implementations are proposed for naive bayes algorithms yet they come at the cost of hardware requirements [236]. [237] proposes a solution to solve the underflow and overfitting problems in big data. The method uses a robust function that works based on average of condition probabilities of all attributes and calculation of dependency of the attributes on the class attribute. Parallel versions of existing discretization methods are also proposed to address the challenge of big data [239]. Yet, more research is required in these open issues.

Support vector machines

SVM is known for it‘s accurate results yet the computational complexity of SVM is quite high on big datasets [240],[241]. Inspite of this computational complexity, SVM uses certain optimization techniques like grid search method for parameter tuning. These optimization techniques are not suited for big datasets [244]. Though certain parameter optimization techniques such as stepwise optimization is proposed in [244] for big datasets, more research is needed in this area. Solutions such as implementing SVM on a quantum computer [240] to reduce it‘s time complexity is proposed. Again, they come at the cost of hardware. Parallel implementation of SVM using mapreduce technique is proposed [241] yet they may end up in local support vectors that may be far away from the global support vectors. [242] proposes a distributed version of SVM where global support vectors are achieved by retaining the first and second order statistics of the big dataset. Though there are many parallel versions of SVM, only a very few parallel tools are available in open source for parallel SVM. Moreover, these tools also require proper tuning [243].

K means clustering

The major problem of k means clustering with big data is it‘s computational complexity as the distance calculation and convergence rate incurs more time with increased number of observations and features. But it can be easily parallelized using big data platforms [245]. Though parallelization is easy with k means clustering using techniques like mapreduce, the I/O and communication cost increases due to repeated reading added with the iteration dependence

property of k means [246]. Methods like subspace clustering and sampling are used to reduce the iteration dependency property of k means [246]. Yet, the choice of correct sampling method and partitioning technique in case of subspace clustering adds to the big data challenges. Indeed, the size of the sample data is more than half the original data in most of the methods and hence the computational complexity still persists [248]. Optimization of initial clusters using techniques like choosing the data points in high density space [247] are proposed. Though they can avoid outliers, they still suffer from the same computational complexity owing to the distance calculation. Dimensionality reduction techniques are also proposed but they come with their own drawback that the cluster in projected space may not comply with the clusters in actual space [248]. Hybrid methods are proposed by combining projection techniques with sampling and few other techniques like visual assessment of cluster tendency. But the research on hybrid techniques are still in it‘s initial state [248].

Hierarchical clustering

Hierarchical clustering also suffers from the drawback of computational complexity and in fact it incurs more time than k means clustering when the size of the dataset is large [252]. Techniques such as building clusters using centroids [250] and usage of cluster seeding [252] are proposed to reduce the computational complexity of hierarchical clustering. Partitioning the sequence space into subspaces using partition tree is also proposed [251] and the clusters are refined in the subspaces. Fast methods to compute the closest pair is also proposed to reduce the computational cost. Yet, these methods are very specific to the particular problem. Moreover, the partitioning techniques and the cluster seeding techniques should be chosen wisely. Visual assessment of tendency are also used to return single linkage partitions on big datasets [249] yet the study of tendency curves have to be clear.

Density based clustering

Density based algorithms are better compared to partitioning algorithms on big data and data streams because it can handle datasets of arbitrary shapes. It is also not required to specify the number of clusters and it can handle noise effectively. But, with the high speed of evolving data streams and high dimensional big data, density based clustering is finding many challenges. Though few methods are found to perform better, they still suffer from open challenges such as too many parameters to set, memory constraints, handling different kinds of data such as categorical , continuous etc [254]. Big data platforms such as hadoop is used for parallelization in density based clustering. Yet, there is a need to choose the shuffling mechanism, partitioning technique and work load balancing efficiently [253]. Moreover, density based clustering algorithms such as OPTICS cannot be parallelized as such and either improvements or new algorithms have to be proposed to handle large datasets [255]. Few enhancements are carried out in OPTICS and other density based clustering algorithms to support parallelism. Yet, they are very specific to the problems they address. For example, [256] uses spatio temporal distance and temporal indexing for parallelization which is more specific to the spatio temporal data and [257] proposes a method that is specific to the electricity data.

Neural networks

The major challenge that neural network faces with big data is the time taken for training phase as large data sets require more training iterations or epochs [260],[261]. As a result, the computing power required becomes high [258] and in turn the energy consumption[260]. Though techniques like mapreduce on hadoop platforms are used [259], the mapreduce design has to be an optimized and efficient one. Hardware solutions such as GPU and memrister are proposed [258], yet they suffer from the major drawback, the cost factor. Optimization algorithms are proposed [259] to optimize the parameters of neural networks. Yet, there is a common perspective that using optimization algorithms increases the computational time due to it‘s convergence property though proper optimization decreases the training time of neural networks inspite of improving the accuracy. Very few researches are carried out in this area for big data with neural networks. The other problems with neural networks on big data include the increase in number of parameters, lack of proper theoretical guideline in the structure of neural networks, insufficient knowledge as only abstraction is extracted, inherent problem in learning etc[261].

Fuzzy systems

Fuzzy systems are of great use in big data due to it‘s ability to handle uncertainty. To cope up with the big data requirements, fuzzy systems are designed that distributes the computing operations using mapreduce technique [61]. But the major problem with mapreduce is the overhead taken to reload the job everytime during it‘s execution. Moreover when there are more number of maps, the imbalance problem has to be dealt with carefully. Spark which has in memory operations and resilent distributed databases is more efficient than mapreduce but unfortunately there are no big works that integrate fuzzy systems with spark [263]. Fuzzy systems are also found to be more scalable as they represent the data as information granules [262]. Yet, good granular techniques in combination with fuzzy classification systems exclusively for big data is required [262]. The fuzzy techniques designed for big data should be tested for real world problems and more general fuzzy techniques need to be developed rather than the techniques designed to address specific problems [262].

Deep learning

Deep learning is used for big data due to it‘s accuracy and automatic learning of hierarchical data representations [264]. Yet, the major problem with deep learning is the requirement of high performance systems. There are other areas that need to be explored in deep learning for big data problems. These include transfer learning with deep architectures, deep online learning that is still in the initial stage of research [264], incremental learning with deep architectures [265], working with temporal data [266] etc. Indeed, constant memory consumption due to the fact that

deep learning is usually performed on very big data that involves millions of parameters and many CPUs is another problem. Hence deep learning requires the use of excessive computational resources. Moreover, deep learning also faces the challenge of determining the number of optimal parameters, learning good semantic data representations as they are known for representing only the abstract detail etc [265]. Lot of research works is required to address the interpretablity problem[266].

Ensemble algorithms

The major problem with ensemble algorithms for big data is it's computational time [268]. As ensemble techniques require the use of different classifiers, the computational time they require is generally high and this increases when the data is big. Ensemble techniques are known for their diversity as they use different kinds of classifiers and aggregate the results. Though many ensemble techniques are developed for static data, there are no big research works carried out in studying the diversity of ensemble techniques on online streaming data. Since the different classifiers used on streaming data already differs in the data they use, a proper study of the advantages of ensemble techniques on streaming data is required. Proper pruning techniques is also an area to be explored [267]. There are few works where ensemble techniques for big data is parallelized with mapreduce [268], yet they are not tried on platforms such as spark that are proved to be more efficient than mapreduce.

4 Predictive analytics on big data across different domains

4.1 Healthcare

Big data is generated by lot of industries and health care is one among them. Huge amount of big data is generated from wearable devices in patients, emergency care units, etc. Structured data such as electronic patient record, health record, unstructured and semi structured data such as scanned images, clinical notes, prescriptions, signals from emergency care units, health data from social media are few examples of big data generated from health care domain. Predictive analytics on health big data helps in predicting the spread of diseases [8], [9], predicting chances of readmission in hospitals [10], predicting the diseases at an early stage, [11], [12],clinical decision support system to identify the right treatment for the affected patients, hospital management etc [13]. A detailed overview about the use of predictive analytics in health informatics is presented in [14]. The paper discusses about the applications of big data predictive analytics in health informatics, techniques for the same, opportunities and challenges.

Research gaps

Apart from storage, processing and aggregating different types of data in health care, identifying the dependencies among different data types is still an open challenge that requires optimal solution. Another challenge is with data compression methods. Though various methods are available for big data compression like FPGA, lossy image compression, they might not suit well for medical big data since medical images shouldn't lose any information [44]. An another area of improvement is in predicting the spread of diseases earlier. Though there are certain works carried out in this area, few important features were not taken into consideration for prediction. For example, though environmental attributes are used as input for predicting the spread of disease, certain inputs related to biological and sociobehavioral is excluded in the proposed approach in [8]. Proper dimensionality reduction techniques and feature selection are not considered in few works. Merely knowledge of domain experts are used for feature selection in health big data [10]. Clinical decision support systems is another challenge to work with. Though clinical decision support systems are developed, the success rate is very less. The decision support system should be developed considering both the patient's and physician's perspectives as patients' acceptance is very important for these systems. It can be of greater importance for emergency care units. Few challenges to work on include privacy issues, proper training for clinicians, quality of data etc. Such systems help in taking precautionary actions like identifying low risk patients, work on hospital readmission rates etc [60]. Radiation oncology is another area open to researchers. Building an integrative model for radiation oncology to be used as decision support system will be of great help for clinicians [10]. More concentration on genomic analysis is required since the present applications of clinical prediction uses genomic data. Research on functional path analysis of genomic data has to be concentrated upon [44]. Research works on handling noise, complexity, heterogeneity, real time, privacy in clinical big data is the need of the hour [14].

4.2 Education

Education field is another domain generating lot of big data using sensors, mobile devices for applications like learning management system, online assignments, exams, marks, attendance etc. Social media is also widely used by students and instructors as forums [15]. Predictive analytics in data generated from these devices and institutions help to predict the effectiveness of a course [15], learning method [16],[17], student's performance [18], [20], institution's performance [21] etc. Such predictions also help to personalize instructions by customizing the learning experience to each student's skills and abilities. [19] discusses about big data opportunities and challenges in education field. [17] and [21] also discusses about the scope of predictive analytics in educational big data.

Research gaps

Firstly, there are very few works carried out for predictive analytics in education field. There are very few papers in this field and most of them are review and survey papers. Hence researchers can focus on this field. The major

challenge involved in this field is social and ethical challenges. Since the student‘s and institution‘s individual data are used for prediction purposes, many students and institutions might not want their data to be exposed [68]. In the recent days, many institutions allow students to bring their own devices for language learning. Hence massive amounts of data is generated and scalability is an important area to concentrate in future [19]. Another area to work on is in integrating data from different sources. Student data are available in multiple sources like social media, schools, district offices, universities etc. Few are structured and few unstructured. A more focus on this area can help prediction.

4.3 Supply chain management

Predictive analytics helps in supply chain management. Accenture is a company that has implemented big data solutions for prediction in supply chain management [22]. Prediction in supply chain management helps to improve customer relationships by regular interactions with them thereby helping in understanding their satisfaction level, product recommendations [22], predicting supplier relationships [23], reduce the customer waiting times [24], improve the production based on demand [25], [26], manage the inventory effectively [27] and reduce risks in the process of supply chain [28].[29] discusses about the advantages of predictive big data analytics in this domain. Product development is another area where big data solutions can help the process.

Research gaps

Though there is more scope for prediction in supply chain management, very few industries have implemented it. Probably because of the hesitation to invest and the lack of skillset. Models which can reduce cost can be proposed for supply chain management processes. Hiring data scientists with domain knowledge helps the industries to move towards big data solutions for efficient supply chain management [22]. Though some of the companies use analytics in supply chain management, most of them are ad-hoc and situation specific. Predictive analysis on other areas like improvement in demand driven operations, better customer supplier relationships, optimization of inventory etc can be more concentrated upon [22], [27]. Generalized models for prediction in supply chain industry can be more focussed on. Though [23] proposed a model based on deduction graph, it is not tested on variety of product designs. Privacy of data is not considered. The approach also uses lot of mathematical techniques. Hence approaches using simple techniques can be developed. More solutions for supply chain management considering both strategy and operations has to be focussed upon [26].

4.4 Product development and marketing industry

[30] presents a white paper about the scope of predictive analytics for product development process. Marketing is a part of almost all the sectors and prediction in marketing has gained more importance because of its direct impact in business and income. Predictions using big data solutions for marketing helps to acquire customers, develop customers and retain customers. Prediction in product development and marketing industry helps to validate the design of the product, predict the demand and supply thereby increasing the sales and improving the customer experience.

Research gaps

There is no single technology available to address all the big data requirements. Big data solutions have to be integrated with other approaches and techniques to support predictive analytics. Researchers can concentrate to work on a single technology that can address all the requirements [30]. Also, different processes have to use different techniques and approaches specifically designed for them. For example,semiconductor manufacturing process should consider the spatial, temporal and hierarchical properties in manufacturing process as the existing algorithms doesn‘t suit well for them. Specific solutions can be proposed for different product development industries [55]. More work on implementing the machine learning algorithms in different areas of marketing and integrating them together can be a scope of work for researchers [45].

4.5 Transportation

‘Smart city’ is a diction that is of common talk in today‘s world. A smart city uses the information collected from sensors operating over the cities to help in the administration of the cities. Many research works are carried out in this area. Intelligent transportation systems is required for building smart city. Sensors generate lot of information that require big data solutions for processing and prediction. Predictive analytics using big data solutions for transportation has lot of applications like predicting the traffic and controlling it efficiently [31], [32], [33],predicting the travel demand and making effective use of the infrastructure thereby reducing the waiting time of the passengers [34], [35], automatic control of traffic signals [36] and predicting the transport mode of a person [37].

Research gaps

With the advent of many devices, lot of information is generated in the field of transportation from various sources. New tools to integrate data from sensors and latest devices to traditional data sources is required. Researchers can focus on developing such tools [34]. The capability of the real time traffic data collection service should be improved since video and image data are all collected [36]. Proper methods to handle correlations among data and uncertainty in data is also required in this field since the data

generated is temporal and spatial in nature for transportation. Readings from sensors are also uncertain [34]. Another area to concentrate is on using other deep learning approaches to predict traffic flow for better performance. The prediction layer used in [33] is just logistic regression. More powerful predictors can improve performance. Data fusion for social transportation data is still in a preliminary stage. GPS from taxi driver only give information about origin and destination but not on travel demand. Mobile data are used for travel demand but can't estimate travel time on roads. Hence a proper fusion approach has to be used to integrate data from several sources.Web based agent technology for transportation management and control is a research direction [35]. Software robots to monitor the state of drivers, check the condition of cars, evaluate safety environment is a research in progress [35]. There is more scope in predicting the transport mode of a person. [37] used only sensor information for classification whereas future work can concentrate on integrating information from cloud too. Advanced techniques to remove noise and outliers can be worked upon.

4.6 Other domain areas

Agriculture can be benefited using predictive analytics. Now-a-days sensors and UAVs finds its use in agriculture. Sensors are used to find the effectiveness of certain type of seed and fertilizer in different places of the farmland. Big data solutions are used to store and analyze this information to improve the operations in agriculture. Additional information like predicting the effect of certain diseases on crops helps to take precautionary measures. Farmers do predictive analytics in agricultural big data to lower costs and increase yields. The use of different fertilizers, pesticides is used to predict the environmental effects [38], [39]. Big data prediction finds its application in banking. Analysis in browsing data helps to acquire customers. Defaulters are predicted by mining Facebook data. Banks use sentiment analysis to analyze the customer needs and preferences and motivate them to buy more products thereby reducing the customer churn. Facebook interactions, tweets, customer bank visits, logs, website interactions are used as sources for sentiment analysis. A 360 degree view of customer interactions is analyzed to prevent churning of customers. Certain features like account balance, time since last transaction all helps banks to frame rules and identify the customers who are about to churn. Big data prediction helps to identify hidden customer patterns. Large sample of outliers are analyzed to predict fraud detection in banks [40]. There are also few works carried out for prediction in library big data. Libraries work with online journals and resources which are again voluminous. Predictive analytics is used in library big data for useful extractions related to learning analytics, research performance analytics etc. The user search behavior, log behavior are all analyzed to extract useful information. An other industry where predictive analytics finds its importance is telecommunication industry. Lot of applications today like Whatsapp uses different kinds of data that include structured, unstructured and semi structured. Predictive analytics in telecommunication industry focusses mainly on customer satisfaction [41]. Predictive analytics in big data helps business too. Big data analytics platforms of different providers help in personalization. The extent to which they support personalization differs in different platforms. Many big data platforms like KNIME, IBM Watson analytics are all finding its use in personalization [42]. Prediction in big data is used extensively in robotics field also. Robots communicate with many other systems. Sharing of information between robots and smart environments, comparing the information the robot has with other systems improves the robot intelligence [43]. Movie industry uses big data solutions for predicting the success using social media. Enormous data gets accumulated in social media like Wikipedia, Facebook and Twitter. Self-aware machines are finding its way in industries with the help of big data and cloud computing techniques. These machines are capable of monitoring their health condition on their own and take smart decisions on their maintenance.

5 Comprehensive challenges

Big data has its own challenges in terms of storage, processing, analytics etc. We restrict this paper to address the overall challenges involved in predictive analytics on big data and to throw light on few techniques used and state of the art work done in handling these challenges. The overall challenges are categorized under six headings shown in figure 3.

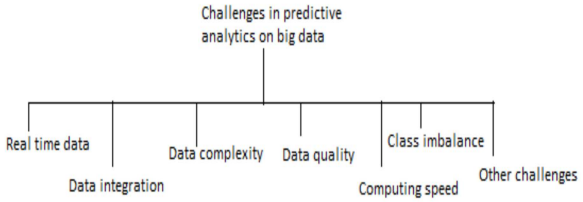


Figure 3: Comprehensive challenges of Predictive analytics on big data - taxonomy

5.1 Real-time data

Handling real time data is one among the major challenges in predictive analytics on big data. Few predictions such as predicting the early outbreak of the disease to take care of public health [9], real time recommendation system for marketing requires real time data from social sites to be collected.

Firstly, Latency is one of the main parameter to be taken care of when working with such data. Secondly, techniques

to handle interactive queries is important during predictions with these data. Thirdly, predictive algorithms should be integrated with solutions handling real time data for effective prediction.

Some big data solutions and machine learning algorithms are used in handling the above challenges with real time data. [44] states that spark and storm helps in collecting the data without latency. Hadoop platform are used to carry out Extract, transform, load(ETL) operations. Hbase, HiveQL are used to work on interactive queries. Open source software applications like Cassandra, MongoDB are used to achieve scalability and performance. Apache Mahout Machine learning library is used to run predictive algorithms on top of Hadoop [30]. Yet these techniques are generalized and their performance differs depending on the nature of the data. [72] proposes a task level adaptive mapreduce framework for real time streaming data in health care applications. The scaling capability is designed at task level that improves scalability in real time streaming data from sensors. It is proved to cope up with the velocity of the streaming data. But this approach was tested only on health care applications.

Few traditional data mining algorithms are also proposed for real time data. [44] states that algorithms such as Naive bayes can be used for sentiment analytics to extract the words from the twitter or other sites. Logistic regression, nearest neighbours are used for customer segmentation and to predict the probability that the customer will click the advertisement. [73] uses naive bayes to extract information from tweets texts. [4] proposes the use of Extreme learning machine to enhance the speed for processing real time data. [30] uses random forests and bayesian techniques to predict the crash and congestion in real time traffic monitoring. Yet, when the velocity at which the real time data enters increases, the performance of these traditional data mining algorithms deteriorate.

Mobile computing techniques and cloud infrastructure are used with big data platforms and data mining algorithms to handle real time data. [46] proposes a monitoring platform Context-Aware platform using integrated mobile services(CAPIM) to make the life of smart city easier. CAPIM collects the general traffic information, stores in the mobile device and uploads when the wi-fi is available. Drivers are provided feedback about the traffic information that helps to take decisions. More visualization output is presented to the users using google maps and the services in turn send data about his locality on his social accounts like twitter etc. [31] proposes the use of techniques like Hadoop, Hadoop Distributed File System(HDFS) and HBase to store the traffic related information. This paper proposes Real time traffic information(RTTI) system for collecting and integrating information from various sources. Massive traffic data sets are utilized transparently with the aid of cloud infrastructure. Cloud and big data is integrated in traffic flow algorithms. The usage of cloud for massive storage and the use of mapreduce techniques and Hadoop HDFS improves the performance of data mining algorithms in predictions with real time traffic flow information. Cloud services like Watson analytics is also used to analyze real time data from social media. Yet there are not many platforms that integrates cloud services, big data solutions and mobile computing. There is a need for data fusion approaches.

Few other related works for real time data include [4] that uses Representative streaming processing systems for processing real time streaming data and [46] that uses Very Fast Decision Tree (VFDT) algorithm and IBM Infosphere streams to analyze the real time streaming population data to predict the spread of cardio respiratory spells. [48] also proposes a multi-dimensional fusion technique that works on Hadoop platform with both real time and offline data. This model suits well for satellite applications where real time data is captured and forwarded to the ground station for refining and prediction. Energy efficient memrister based neural network is used for big data analytics. GPUs are used instead of CPUs to improve the speed up. Recurrent neural networks are used since they prove to be effective for non- linear sequential processing tasks like speech recognition, natural language processing, video indexing etc. [69] uses convolutional neural networks with GPU capabilities to detect real time objects. [70] combines real time measurements from real time databases with static data and uses simple extrapolation technique for prediction in substation automation. [71] proposes a compression technique for real time analog signals. It can handle the big data in real time instruments and optical communications.

5.2 Data integration

This section discusses about the challenges involved in integrating data of different types for prediction. Heterogenous data is one of the main characteristics of big data. Few predictive works require data from different sources to be integrated with the existing data. For example, predictions in educational sector collects data from multiple sources like social media, schools, district offices, universities etc. Few are structured and few are unstructured. [20] proposes a model that integrates information from social media and predicts student involvement and success. Since social media is used by students to share their ideas, feedbacks about courses, sentiment analysis can be used to solve problems by analyzing the most common feedbacks, ideas etc [16]. Healthcare sector also requires integration of information from different sources. [49] uses unstructured data from Google Flu trends to predict the spread of influenza by predicting the region that are most likely to be affected. Past data is combined with real time data for prediction. [11] proposes a predictive model to predict Systemic Lupus Erythematosus, a disease that affects multiple organs by integrating structured data like electronic health records, unstructured and semi structured data like imaging and scan tests(MRI, CT, Ultrasound scan, X-ray), complete blood count, urinalysis. [8] develops a spatial data model to predict influenza epidemic in Vellore, India. Large repos-

itories of data are collected. The developed spatial model is dependent on geographically weighted regression technique. It involves bringing together several data sources like surveillance systems, sentinel data etc to predict the spread of epidemics. Movie industry uses big data solutions for predicting the success using social media. Enormous data gets accumulated in social media like Wikipedia, Facebook and Twitter. [50] proposed a predictive model for predicting the movie box office success in Wikipedia. The major issue with heterogeneous data is that since they may not exactly be the same, there are possible chances that the machine learning results may be affected.

Big data platforms with simulation help to bring together data from different sources and predict the hidden patterns in them. [10] develops a predictive model using Mahout's machine library that works on top of Hadoop with Mapreduce technology to find out the chances of readmission after discharge of persons with heart failure. Data is collected from various sources and integrated using mahout. Mahout machine library is used for prediction that runs mapreduce jobs on Hadoop. The model uses HiveQL for distributed query. Data extraction and integration is done using Hadoop, Hive and Cassandra. Random forest and logistic regression is used in predicting the readmissions. Big data solutions gave good results in terms of time efficiency and scalability. [23] proposes a model for prediction in supply chain management process that uses deduction graph model to visually link competent sets from many data sources both structured and unstructured. Customer preferences and new product ideas are predicted through social media using recent shopping history. Customer response time is improved. Big data solutions such as Apache Mahout is used for machine learning, tableau is used for visualization, Ionosphere for data mining etc. Infrastructure proposed by combining deduction graph model with data mining, proves to provide better results in supply chain management with respect to usability, feasibility etc. [51] proposes the use Mapreduce technology in distributed data management and scheduling for heterogenous environment. The paper proposes a system for social transportation. Statistical approaches, data mining algorithms and visualization analytics are used according to the type of data. It is implemented in hadoop platform but used with other frameworks also like cascading, sailfish, Disco, Skynet, Filemap, Themis etc. High level languages like PigLatin, HiveQL are used for the technology. Hence, very few tools and techniques are available to integrate data from different sources and the advent of devices like sensors and RFIDs kindles the requirement for new tools and techniques.

Oncology can help integration easier in healthcare sector. Radiation oncology ontology is a key component used in data collecting system for better interpretability. Radiation oncology requires aggregation of input from many origins like scans, images and EHR of patients. [52] carried out a survey that explains about the state of art and future prospects of using machine learning algorithms and big data in radiation oncology. [9] states that building an integrative model for radiation oncology to be used as decision support system will be of great help for clinicians. But the research works in oncology is in its very initial stage.

Data warehousing is another concept in which data are integrated from heterogeneous sources. Few examples include [275] that proposes a dimensional warehouse for integrating data from clinical trials, [276] that proposes an architecture for a data warehouse model to integrate health data from different sources etc. The major drawback with data warehousing with big data is that technologies like hadoop, mapreduce etc. has to be integrated with the data warehouse to support the big data requirements. Moreover the ETL operations have to be designed in the architecture to suit big data requirements. There is also no standard architectural framework to design data warehouse for big data and hence the existing architectures designed to suit particular problems lacks flexibility [277]. Concepts such as data lakes are also of use as they are non relational approaches to integrate different types of data from heterogeneous resources. They postpone data mapping until query time. But, the query and reporting capabilities of data lakes are still emerging. They are not as powerful as SQL on relational databases [275]. Few research works such as [278] are proposed that manages the metadata effectively in data lakes to address the big data requirements.

Few related works on data integration for big data includes [74] for their work on a new resource 'Diseases' that aims to find the associations between genes and diseases by integrating automated text mining with existing datasets, [75] for their data integration method 'Optique' based on ontology that integrates streaming and static data as an abstraction layer, [76] for their work on API centric linked data integration that discusses about the use of API to extend the classical linked data application architecture to facilitate data integration, [78] for their work to propose a new approach for integration based on semantics. This approach converts the data sources in different formats to nested relational models initially and then imports only a subset of large datasets to build the model thereby coping up with the data size problem. A review of data integration in life sciences along with its challenges are discussed in [77]. [79] proposes the use of fused lasso approach in regression coefficients to learn heterogeneity when combining data from heterogeneous sources.

5.3 Data complexity

Large complex datasets such as genomic datasets have to be dealt in few predictive tasks. Such complex datasets make the predictive task difficult. Firstly, storage and processing of such complex data becomes a problem. Secondly, understanding such complex data to build predictive models need to be taken care of. Thirdly, either enhancements in existing data mining algorithms or new data mining algorithms have to be proposed to handle such complex datasets.

[44] states that storage techniques like HDFS, apache

Hadoop helps in storing and processing the big data effectively. Using distributed platform like mapreduce prevents the over utilization of the resources. A detailed survey on map reduce technology is discussed in [6]. Cloud platforms also help in handling complex data. Cloud platforms in health care like "PCS-on-demand"are found to be effective in storing and sharing of healthcare information with its cloud infrastructure. Mapreduce is used to parallelize the processing. Mahout library is used for machine learning algorithm to process the images, signals and genomic data. MongoDB is used for storage because of its high availability, performance and easy scalability [45]. [25] proposes the use of Microsoft Azure, a cloud platform for data storage in inventory management for supply chain process. Data sources for inventory management are internal like RFID, sensors etc. They generate lot of data and big data solutions help in processing them efficiently. NoSQL is used for data access. Batch analytics is done using Apache Hadoop. [32] proposes a model that sends driving guidance to vehicles with cloud computing technique incorporated to big data. Big data solutions [53] are used for spatial data analytics and Cloud solutions in big data platforms are used for predictive analytics in tactical big data [54].

Visualization analytics and clustering helps in understanding complex datasets. [12] develops a predictive model for diabetic data analysis in big data. Association rule mining is used to find association between the laboratory results and diabetes type of the patient. Clustering of similar patterns and classification of health risk value by patients health condition is done and predefined deductive rules are derived to predict the diabetes. The predictive model uses Hadoop/mapreduce environment for parallel processing. Visualization also helps in predicting hidden insights from the data. [13] proposes a model for hospital management. The temporal information helps to understand the clinical activities. Proper visualization and clustering of this temporal data helps to understand the abnormalities. [55] proposes an optimization framework for wafer quality prediction in semiconductor manufacturing process that uses clustering to identify similar behavior pattern over time for chambers. [56] presents a survey on clustering time series data. Abnormality can also be discovered thereby helping quality control and fault diagnostics. But since time series clustering is mostly for unstructured data, a co-clustering pattern is formulated for this problem with constraints to match the tools and the chambers. Visualizing the data effectively helps in prediction. Aggregation and multi-dimensional analysis is also used in big data to extract knowledge from them. AsterixDB, DGFIndex are used that helps in aggregation and multi-dimensional analysis for big data [57]. Yet, new frameworks for visualization techniques and multi-dimensional analysis need to be developed exclusively for big data.

Few data mining algorithms are used in complex datasets such as genomic data and signals. [47] uses Nearest Centroid Based Classifier (NCBC) to predict clinical outcome related to cancer using gene expression data.[58] uses Multipartiate graph for prediction in genomics big data. Deep learning also helps in handling complex data. Lot of research works are carried out in clinical image processing with deep learning. [33] uses deep learning for predicting traffic flow. A stacked autoencoder model trained greedily learns traffic flow features. It uses spatial and temporal correlations.The medical data including cardiac MRI involves signal processing. There are many statistical learning tools for signal processing. [59] uses signal processing techniques such as kernel based interpolators and timely matrix decompositions for big data. Yet, more research works are required for signal processing with big data and on genomic data analysis.

Few related works to handle complex data include [80] for their work on anytime density based clustering for complex data to improve scalability, [81] for their work on interactive data visualization to understand complex data, [83] for their work to propose a Pairwise weighted ensemble clustering algorithm to cluster complex data for better understanding, [82] for their work to address scalability problem in complex data by proposing two suboptimal algorithms to address casting complex problem of L1 Norm principal component analysis, [84] for their work to develop complexheatmap package that helps in visualizing and revealing the patterns in high dimensional genomic data.

5.4 Data quality

Low quality data is another challenge in predictive analytics. The source data in some applications like from emergency care, sensors may be of poor quality. Predictive analytics techniques on such low quality data need some sophisticated techniques to be applied on the existing algorithms. Low quality data may also be due to the fact that some predictive works fail to consider important attributes required for prediction. For example, [8] predicts the spread of the disease using environmental attributes but fail to consider the biological and socio-behavioral attributes. Data may also be incomplete and inconsistent in certain cases.

Generally techniques like Mathematical or logical regression might work for low quality data [60]. [4] states that advanced deep learning methods, statistical learning theory of sparse matrix are used to overcome the challenge of incomplete and inconsistent data. Techniques like Watson analytics are used to overcome the challenge of low value density data. More works on identifying proper correlations among inconsistent data is required.

Proper dimensionality reduction techniques and feature selection also help to improve the data quality. [9] states that merely knowledge of domain experts are used for feature selection in most of the predictive works.

The nature of data differs depending on the application. For example,semiconductor manufacturing process should consider the spatial, temporal and hierarchical properties in manufacturing process as the existing algorithms doesn't

suit well for them. Specific solutions can be proposed for different domains depending on the nature of the data [54].

Few related works on predictive analytics with low quality data include [85] that proposes an extension to likelihood method to handle low quality data, [86] to propose a method that uses data mining tasks such as clustering to extract patterns from noisy data in market segmentation, [87] to propose a new algorithm based on C4.5 decision tree that uses imprecise probabilities in classifying noisy data, [88] that proposes a new algorithm for extreme machine learning to work efficiently in the presence of outliers. [89] proposes a hybrid feature selection scheme to reduce the performance deterioration caused by outliers in predictive analytics.

5.5 Computing speed

Computing speed is one of the important challenges to be handled during predictions on big data. Most of the wearable devices consume more power and the algorithms used on them are computationally intensive. Computing speed of predictive algorithms on big data also increases due to its volume.

Parallel computing techniques help to overcome the challenge with respect to volume and computing speed. [4] proposes the use of alternating direction method of multipliers to overcome the challenge with respect to volume since it acts as a platform for distributed frameworks with parallel computing. Mapreduce is used to work parallel on the chunks of the big data. [63] proposes a data mining algorithm K Nearest neighbor based on Spark(KNN-IS) for classification in big data. The algorithm uses Mapreduce technology for parallel processing of the training data set. Though Hadoop works well with mapreduce, it has its own limitations like latency which is overcome by spark's in-memory computations. Map reduce is used on spark for KNN to yield better results in terms of time and accuracy. Resilient distributed databases are used on spark platform. Sometimes medium quality predictions with low latency perform better than high quality predictions with more latency. [90] proposes parallel random forest algorithm in spark cloud computing platform to improve the computational efficiency of big data analytics. A parallel version of deep neural network training is proposed in [91].

Feature selection techniques like Representation learning, deep learning and dimensionality reduction are also used to reduce the computing speed since the unnecessary features are eradicated. Yet, the computational complexity of certain feature selection techniques like wrapper approaches is high and researchers are working on it. [92] proposes a hierarchical attribute reduction algorithm using mapreduce in which attribute reduction process is executed in parallel. [93] proposes fast minimum redundancy maximum relevance algorithm for feature selection in high dimensional data.

Fuzzy techniques aid in reducing the processing time. Researchers worked towards reducing the time for processing using fuzzy rules on data with lot of input features. An algorithm named Chi-Fuzzy rule based classification systems(Chi-FRBCS) is proposed. It works on Mapreduce framework and uses linguistic fuzzy rules. Two versions of Chi-FRBCA algorithms are proposed - Chi-FRBCS BigData-Max and Chi-FRBCS BigData-Ave. Experiments are conducted on six different big data problems set and Chi-FRBCS is proved to be effective in terms of processing time and accuracy [61]. [62] proposed a big data algorithm called FMM based on fuzzy rules for sentiment analysis in social media. A parallelized algorithm FMM with mapreduce is also used and that proves to be effective in terms of accuracy compared to the other techniques. The algorithm is made to work on twitter data and is observed that the execution time is much lesser for big data.

[64] states that the hardware solution is effective in terms of energy savings, power efficiency. Scientific applications use multi-dimensional data sets. Processing has to be faster compared with other applications because of the velocity at which it arrives. For example, predicting climate change requires fast processing. [65] proposes the use of an I/O in-memory server for scientific big data analytics applications. [37] proposes SVM polynomial degree 3 kernel to reduce the computational complexity and memory requirement during classification. This model detects the transport mode of a person whether he or she is walking, jogging or going in bike. Hardware changes are also done by using virtual gyroscope, accelerometer and few other hardware devices to ensure that low power consuming devices are used. The memory consumed by the algorithm is less.

5.6 Class imbalance

Class imbalance is another problem in certain predictive works. Techniques like oversampling and undersampling are used in class imbalance problems.

Some machine learning algorithms are effective in solving class imbalance problems. [66] proposes an algorithm ROSEFW-RF for contact map prediction. It is a classification task related to protein structure where there are very few positive samples available. This algorithm is based on key-value pairs and uses mapreduce approach for distributed processing. Predictive model is constructed using random forest. The class distribution is balanced through random oversampling. Irrelevant features are removed through feature weighing. Oversampling is found to be more robust than undersampling and cost sensitive learning when number of maps is increased in mapreduce technique. The test data is classified. Experiments are conducted in bioinformatics data and ROSEFW-RF algorithm is the winner algorithm for imbalance big data classification problem. [67] proposes Random forest with Mapreduce for prediction on imbalanced big data. Five different versions of random forest algorithms are used in imbalanced big data classification with Mapreduce approach. - RF - BigData, RF-BigDataCS, ROS+RF-BigData, RUS+RF-

BigData, SMOTE+RF-BigData. Random forest techniques is found to work well for imbalanced big data classification.

Few other related works on class imbalance problem include [94] that proposes an ensemble method to handle both online learning and imbalance learning using oversampling and undersampling bagging techniques, [95] that proposes a diversity based oversampling approach to create new instances for minority class, [96] that proposes a new ensemble method 'hardensemble' to handle class imbalance, [97] that uses extreme learning machine to handle class imbalance problem both at feature and algorithmic levels.

Class imbalance problem is another area that require researcher's attention. A mixed strategy of oversampling and undersampling can be tried to boost performance [66].

5.7 Other challenges

Few other challenges include privacy issues, lack of proper training to the domain experts, social and ethical challenges etc. For example, the decision support system in healthcare should be developed considering both the patient's and physician's perspectives as patients' acceptance is very important for these systems. It can be of greater importance for emergency care units. Privacy issues, proper training for clinicians, quality of data are all few issues to be considered in this case [60]. In education sector, since the student‘s and institution‘s individual data are used for prediction purposes, many students and institutions might not want their data to be exposed [68]. Hiring data scientists with domain knowledge helps the industries to move towards big data solutions for efficient supply chain management [22].

6 Potential research directions

From the above discussions on predictive analytics with big data, it has been observed that this field is readily open for researchers. The potential research directions are summarized.

6.1 Data management

- Since real time data includes the collection of lot of video and image data in the present scenario, there is a need to improve the real time data collection service. Researchers can work on the same.
- Developing framework fusion approaches to integrate big data solutions, cloud computing techniques and mobile computing techniques is another area for research as there is no single technology available to address all the big data requirements. Researchers can concentrate to work on a single technology that can address all the requirements to support predictive analytics.
- With the advent of many devices, lot of information is generated in different domains from various sources. New tools to integrate data from latest devices to existing data sources is required. Researchers can focus on developing such tools.
- Data partitioning methods, indexing and multidimensional analysis on big data are few other topics for researchers.
- Scalability is an important area to concentrate in future as massive amounts of data is generated.

6.2 Algorithms and solutions

- Enhancements in traditional data mining algorithms to handle and analyze real time data or developing new algorithms for the same is another promising research area available for researchers.
- Identifying the dependencies and semantic features among heterogeneous data types is still an open challenge requiring favorable solutions due to the biased view of data distribution.
- New visualization techniques and frameworks can be developed for effective interpretation of complex data.
- Genomic data analytics is in its initial stage. Works such as functional path analysis on genomic data is an area open for researchers.
- Data compression methods is also a challenge. Big data compression methods like FPGA, lossy image compression, might not suit well for certain types of big data like medical images as they shouldn‘t lose any information. Hence there is a need for new data compression methods exclusively for specific big data types.
- Oncology, semantic web are all areas to be concentrated in machine learning for big data.
- Establishing correlations among uncertain data, temporal and spatial data is another area to work on for researchers.
- Overfitting still remains as an open issue and researchers can focus on developing better solutions without much compromise on accuracy and cost.
- Using ensemble techniques in mapreduce platform is another area that can be concentrated on to improve accuracy
- Specific solutions can be proposed for different domains depending on the nature of the data.
- Class imbalance problem is another area that require researcher's attention. A mixed strategy of oversampling and undersampling can be tried to boost performance

7 Conclusion

The paper provides an overview of core predictive models and their challenges on big data. We discussed the scope of predictive analytics on big data generated across different domains and few research gaps are identified. Though the mathematical approaches may not suit well for big data, we found that the data mining approaches and machine learning techniques used for prediction have their base from the mathematical approaches. The choice of predictive model depends on the nature of application and data in hand. Finally we presented comprehensive challenges of predictive analytics on big data and state of the art techniques used to address the challenges. Based on our discussion, we also presented a separate section on future directions for research.

References

[1] Chiang, Roger H.L and Goes, Paulo and Stohr, Edward A (2012) Business intelligence and analytics education, and program development: A unique opportunity for the information systems discipline, *ACM Transactions on Management Information Systems (TMIS)*, https://doi.org/10.1145/2361256.2361257.

[2] Jain, A. K. and Murty, M. N. and Flynn, P. J (1999) Data Clustering: A Review, *ACM Comput. Surv.*, 264–323, https://doi.org/10.1145/331499.331504.

[3] Benjamins, Richard.V (2014) Big Data: from Hype to Reality?, *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14), ACM*, https://doi.org/10.1145/2611040.2611042.

[4] Qiu, Junfei and Wu, Qihui and Xu, Yuhua and Feng, Shuo (2016) A survey of machine learning for big data processing, *EURASIP Journal on Advances in Signal Processing*, https://doi.org/10.1186/s13634-016-0382-7.

[5] Chen, Ju-Chin and Liu, Chao-Feng (2015) Visual-based Deep Learning for Clothing from Large Database, *Proceedings of the ASE BigData and SocialInformatics, ACM*.

[6] Sakr, Sherif and Liu, Anna and Fayoumi, Ayman G (2013) The Family of Mapreduce and Large-scale Data Processing Systems, *ACM Comput. Surv., ACM*, https://doi.org/10.1201/b17112-3.

[7] Hung, San-Chuan and Kuo, Tsung-Ting and Lin, Shou-De (2015) Novel Topic Diffusion Prediction Using Latent Semantic and User Behavior,*Proceedings of the ASE BigData and SocialInformatics, ACM*.

[8] D. Lopez and M. Gunasekaran and B. S. Murugan and H. Kaur and K. M. Abbas (2014) Spatial big data analytics of influenza epidemic in Vellore, India,*IEEE International Conference on Big Data*, https://doi.org/10.1109/BigData.2014.7004422.

[9] Curran, Martina and Howley, Enda and Duggan, Jim (2016) An Analytics Framework to Support Surge Capacity Planning for Emerging Epidemics,*Proceedings of the 6th International Conference on Digital Health Conference, ACM*, https://doi.org/10.1145/2896338.2896354.

[10] Zolfaghar, Kiyana and Meadem, Naren and Teredesai, Ankur and Roy, Senjuti Basu and Chin, Si-Chi and Muckian, Brian (2013) Big data solutions for predicting risk-of-readmission for congestive heart failure patients,*IEEE International Conference on Big Data*, https://doi.org/10.1109/BigData.2013.6691760.

[11] S. Gomathi and V. Narayani (2015) Implementing Big Data analytics to predict Systemic Lupus Erythematosus,*International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, https://doi.org/10.1109/ICIIECS.2015.7192893.

[12] Eswari, T and Sampath, P and Lavanya, S and others (2015) Predictive Methodology for Diabetic Data Analysis in Big Data,*Procedia Computer Science*, pp:203–208, https://doi.org/10.1016/j.procs.2015.04.069

[13] Tsumoto, Shusaku and Hirano, Shoji (2015) Analytics for Hospital Management,*Proceedings of the ASE BigData and SocialInformatics, ACM*.

[14] Fang, Ruogu and Pouyanfar, Samira and Yang, Yimin and Chen, Shu-Ching and Iyengar, S.S (2016) Computational health informatics in the big data age: a survey,*ACM Computing Surveys (CSUR)*, pp.12, https://doi.org/10.1145/2932707.

[15] The center for Digital education (2015) : Big data in education report. Technical Report, https://blog.stcloudstate.edu.

[16] Sin, Katrina and Muthu, Loganathan (2015) Application of big data in education data mining and learning analytics?A literature review,*ICTACT Journal on Soft Computing*, pp.1–035, https://doi.org/10.21917/ijsc.2015.0145.

[17] Oracle (2015) Big data education. Technical Report, www.oracle.com.

[18] Jo, Il-Hyun and Kim, Dongho and Yoon, Meehyun (2014) Analyzing the Log Patterns of Adult Learners in LMS Using Learning Analytics,*Proceedings of the Fourth International Conference on Learning Analytics And Knowledge, ACM*, pp.183–187, https://doi.org/10.1145/2567574.2567616.

[19] Wang, Yinying (2016) Big opportunities and big concerns of big data in education,*TechTrends, Springer*, pp.1–4 , https://doi.org/10.1007/s11528-016-0072-1.

[20] Niemi, Gitin and David, Elena (2012) Using Big Data to Predict Student Dropouts: Technology Affordances for Research,*International Association for Development of the Information Society*.

[21] Kellen, V and Consortium, Cutter and Recktenwald, A and Burr, S (2013) Applying big data in higher education: A case study,*Data Insight and Social BI*.

[22] Accenture (2014) Accenture-Global-Operations-Megatrends-Study-Big-Data-Analytics. Technical Report, https://acnprod.accenture.com.

[23] Tan, Kim Hua and Zhan, YuanZhu and Ji, Guojun and Ye, Fei and Chang, Chingter (2015) Harvesting big data to enhance supply chain innovation capabilities: An analytic infrastructure based on deduction graph,*International Journal of Production Economics*, pp:223–233, https://doi.org/10.1016/j.ijpe.2014.12.034.

[24] Rozados,Ivan Varela and Tjahono, Benny (2014) Big Data Analytics in Supply Chain Management: Trends and Related Research,*6th International Conference on Operations and Supply Chain Management*.

[25] J.Leveling and M. Edelbrock and B. Otto (2014) Big data analytics for supply chain management,*IEEE International Conference on Industrial Engineering and Engineering Management*, pp:918-922, https://doi.org/10.1109/IEEM.2014.7058772.

[26] Wang, Gang and Gunasekaran, Angappa and Ngai, Eric WT and Papadopoulos, Thanos (2016) Big data analytics in logistics and supply chain management: Certain investigations for research and applications,*International Journal of Production Economics, Elsevier*, pp:98–110, https://doi.org/10.1016/j.ijpe.2016.03.014.

[27] Waller, Matthew A and Fawcett, Stanley E (2013) Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management,*Journal of Business Logistics*, pp:77–84, https://doi.org/10.1111/jbl.12010.

[28] Wilschut, Tim and Adan, Ivo JBF and Stokkermans, Joep (2014) Big data in daily manufacturing operations,*Proceedings of the Winter Simulation Conference*, pp:2364–2375, https://doi.org/10.1109/WSC.2014.7020080.

[29] He, Miao and Ji, Hao and Wang, Qinhua and Ren, Changrui and Lougee, Robin (2014) Big data fueled process management of supply risks: sensing, prediction, evaluation and mitigation,*Proceedings of the Winter Simulation Conference*, pp:1005–1013, https://doi.org/10.1109/WSC.2014.7019960.

[30] Intel (2013) Predictive analytics and interactive queries on big data.Technical Report, https://software.intel.com.

[31] Shi, Qi and Abdel-Aty, Mohamed (2015) Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways,*Transportation Research Part C: Emerging Technologies*, pp:380–394, https://doi.org/10.1016/j.trc.2015.02.022.

[32] Yu, Jianjun and Jiang, Fuchun and Zhu, Tongyu (2013) RTIC-C: a big data system for massive traffic information mining,*International Conference on Cloud Computing and Big Data (CloudCom-Asia), IEEE*, pp:395–402.

[33] Y. Lv and Y. Duan and W. Kang and Z. Li and F. Y. Wang (2015) Traffic Flow Prediction With Big Data: A Deep Learning Approach,*IEEE Transactions on Intelligent Transportation Systems*, pp:865-873.

[34] Toole, Jameson L and Colak, Serdar and Sturt, Bradley and Alexander, Lauren P and Evsukoff, Alexandre and Gonzalez, Marta C (2015) The path most traveled: Travel demand estimation using big data resources,*Transportation Research Part C: Emerging Technologies*, pp:162–177, https://doi.org/10.1016/j.trc.2015.04.022.

[35] Yuan, Nicholas Jing and Zheng, Yu and Zhang, Liuhang and Xie, Xing (2013) T-finder: A recommender system for finding passengers and vacant taxis,*IEEE Transactions on Knowledge and Data Engineering*, pp:2390–2403, https://doi.org/10.1109/TKDE.2012.153.

[36] Wang, Chao and Li, Xi and Zhou, Xuehai and Wang, Aili and Nedjah, Nadia (2016) Soft computing in big data intelligent transportation systems,*Applied Soft Computing, Elsevier*, pp:1099–1108, https://doi.org/10.1016/j.asoc.2015.06.006.

[37] Yu, Meng-Chieh and Yu, Tong and Wang, Shao-Chen and Lin, Chih-Jen and Chang, Edward Y (2014) Big data small footprint: the design of a low-power classifier for detecting transportation modes,*Proceedings of the VLDB Endowment, VLDB Endowment*, pp:1429–1440, https://doi.org/10.14778/2733004.2733015.

[38] Stubbs, Megan (2016) Big Data in U.S Agriculture.Technical Report.

[39] Sangwani, G (2016) The Big Future of Big Data,*Business Insider, India*.Technical Report.

[40] Martens (2016) Financial forum.Technical Report, https://www.financialforum.be.

[41] Malaka, Iman and Brown, Irwin (2015) Challenges to the Organisational Adoption of Big Data Analytics: A Case Study in the South African Telecom-

munications Industry,*Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists, ACM*, pp:27:1–27:9, https://doi.org/10.1145/2815782.2815793.

[42] Lopes, claudio and Cabral, Bruno and Bernardino, Jorge (2016) Personalization Using Big Data Analytics Platforms,*Proceedings of the Ninth International Conference on Computer Science & Software Engineering, ACM*, pp:131–132.

[43] Felzmann, Heike and Beyan, Timur and Ryan, Mark and Beyan, Oya (2016) Implementing an Ethical Approach to Big Data Analytics in Assistive Robotics for Elderly with Dementia,*SIGCAS Comput. Soc., ACM*, pp:280–286, https://doi.org/10.1145/2874239.2874279.

[44] Belle, Ashwin and Thiagarajan, Raghuram and Soroushmehr, S.M.Reza and Navidi, Fatemeh and A.Beard, Daniel and Najarian, Kayvan (2015) Big data analytics in healthcare,*BioMed research international*, https://doi.org/10.1155/2015/370194.

[45] EVRY (2014) Big data - white paper.Technical Report, www.evry.com.

[46] Dobre, C and Xhafa, F (2014) Intelligent services for big data science,*Future Generation Computer Systems*, pp:267–281, https://doi.org/10.1016/j.future.2013.07.014.

[47] Herland, Matthew and Khoshgoftaar, Taghi M. and Wald, Randal (2014) A review of data mining using big data in health informatics,*Journal Of Big Data*, pp:1–35, https://doi.org/10.1186/2196-1115-1-2.

[48] Ahmad, Aswais and Paul, Anand and Rathore, Mazhar and Chang, Hangbae (2016) An efficient multidimensional big data fusion approach in machine-to-machine communication,*ACM Transactions on Embedded Computing Systems (TECS)*.

[49] Davidson, Michael W and Haim, Dotan A. and Radin, Jennifer M (2015) Using networks to combine big data and traditional surveillance to improve influenza predictions,*Scientific reports*, https://doi.org/10.1038/srep08154.

[50] Mestyan, Marton and Yasseri, Taha and Kertesz, Janos (2013) Early prediction of movie box office success based on Wikipedia activity big data,*PloS one*, https://doi.org/10.1371/journal.pone.0071226.

[51] Zheng, Xinhu and Chen, Wei and Wang, Pu and Shen, Dayong and Chen, Songhang and Wang, Xiao and Zhang, Qingpeng and Yang, Liuqing (2016) Big data for social transportation, *IEEE Transactions on Intelligent Transportation Systems*, pp:620–630, https://doi.org/10.1109/TITS.2015.2480157.

[52] Bibault,Jean Emmanuel and Giraud, Philippe and Burgun, Anita (2016) Big Data and machine learning in radiation oncology: State of the art and future prospects,*Cancer letters*, https://doi.org/10.1016/j.canlet.2016.05.033.

[53] Chen, Xin and Vo, Haong and Aji, Ablimit and Wang, Fusheng (2014) High performance integrated spatial big data analytics,*Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, ACM*, pp. 11–14.

[54] Savas, Onur and Sagduyu, Yalin and Deng, Julia and Li, Jason (2014) Tactical Big Data Analytics: Challenges, Use Cases, and Solutions,*SIGMETRICS Perform. Eval. Rev., ACM*, pp.86–89, https://doi.org/10.1145/2627534.2627561.

[55] Zhu, Yada and Xiong, Jinjun (2015) Modern Big Data Analytics for Old-fashioned Semiconductor Industry Applications,*Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp.776–780.

[56] Liao, T Warren (2005) Clustering of time series data survey,*Pattern recognition*, pp.1857–1874, https://doi.org/10.1016/j.patcog.2005.01.025.

[57] Cuzzocrea, Alfredo (2015) Aggregation and multidimensional analysis of big data for large-scale scientific applications: models, issues, analytics, and beyond,*Proceedings of the 27th International Conference on Scientific and Statistical Database Management, ACM*, pp. 23, https://doi.org/10.1145/2791347.2791377.

[58] Phillips, Charles A. and Wang, Kai and Bubier, Jason and Baker, Erich J. and Chesler, Elissa J. and Langston, Michael A. (2015) Scalable Multipartite Subgraph Enumeration for Integrative Analysis of Heterogeneous Experimental Functional Genomics Data,*Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics, ACM*, pp.626–633, https://doi.org/10.1145/2808719.2812595.

[59] Slavakis, Konstantinos and Giannakis, Georgios B and Mateos, Gonzalo (2014) Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge,*IEEE Signal Processing Magazine*, pp.18–31, https://doi.org/10.1109/MSP.2014.2327238.

[60] Alexander T. Janke and Daniel L. Overbeek and Keith E. Kocher and Phillip D. Levy (2016) Exploring the Potential of Predictive Analytics and Big Data in Emergency Care,*Annals of Emergency Medicine*, pp.227 - 236, https://doi.org/10.1016/j.annemergmed.2015.06.024.

[61] Victoria Lopez and Sara del Rio and Jose Manuel Benitez and Francisco Herrera (2015) Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data,*Fuzzy Sets and Systems*, pp.5 - 38, https://doi.org/10.1016/j.fss.2014.01.015.

[62] Bing, Li and Chan, Keith C.C (2014) A fuzzy logic approach for opinion mining on large scale twitter data,*Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing* , pp. 652–657.

[63] Jesus Maillo and Sergio Ramirez and Isaac Triguero and Francisco Herrera (2016) kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data,*Knowledge-Based Systems*.

[64] Wang, Yu and Li, Boxun and Luo, Rong and Chen, Yiran and Xu, Ningyi and Yang, Huazhong (2014) Energy efficient neural networks for big data analytics,*Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp:1–2.

[65] Elia, Donatello and Fiore, Sandro and D'Anca, Alessandro and Palazzo, Cosimo and Foster, Ian and Williams, Dean N (2016) An in-memory based framework for scientific data analytics,*Proceedings of the ACM International Conference on Computing Frontiers, ACM*, pp. 424–429.

[66] Triguero, Isaac and del Rio, Sara and Lopez, Victoria and Bacardit, Jaume and Benitez, Jose M and Herrera, Francisco (2015) ROSEFW-RF: the winner algorithm for the ECBDL14 big data competition: an extremely imbalanced big data bioinformatics problem,*Knowledge-Based Systems*, pp.69–79, https://doi.org/10.1016/j.knosys.2015.05.027.

[67] Sara del Rio and Victoria Lopez and Jose Manuel Benitez and Francisco Herrera (2014) On the use of MapReduce for imbalanced big data using Random Forest,*Information Sciences*, pp.112 - 137, https://doi.org/10.1016/j.ins.2014.03.043.

[68] Johnson, Jeffrey Alan (2014) The ethics of big data in higher education,*International Review of Information Ethics*, pp.4–9.

[69] Ren, Shaoqing and He, Kaiming and Girshick, Ross and Sun, Jian (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *Advances in Neural Information Processing Systems 28*, pp.91–99.

[70] S. S. Biswas, A. K. Srivastava and D. Whitehead (2015) A Real-Time Data-Driven Algorithm for Health Diagnosis and Prognosis of a Circuit Breaker Trip Assembly, *IEEE Transactions on Industrial Electronics, vol. 62, no. 6*, pp. 3822-3831, https://doi.org/10.1109/TIE.2014.2362498.

[71] Mohammad H. Asghari and Bahram Jalali (2014) Experimental demonstration of optical real-time data compression, *Applied Physics Letters*.

[72] Fan Zhang, Junwei Cao, Samee U. Khan, Keqin Li, Kai Hwang (2015) A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications, *Future Generation Computer Systems, Volumes 43-44*, pp. 149-160, https://doi.org/10.1016/j.future.2014.06.009.

[73] Yiming Gu, Zhen (Sean) Qian, Feng Chen (2016) From Twitter to detector: Real-time traffic incident detection using social media data,*Transportation Research Part C: Emerging Technologies, Volume 67*, pp. 321-342, https://doi.org/10.1016/j.trc.2016.02.011.

[74] Sune Pletscher-Frankild, Albert Palleja, Kalliopi Tsafou, Janos X. Binder, Lars Juhl Jensen (2015) DISEASES: Text mining and data integration of disease?gene associations,*Methods, Volume 74*, pp. 83-89, https://doi.org/10.1016/j.ymeth.2014.11.020.

[75] Evgeny Kharlamov, Sebastian Brandt, Ernesto Jimenez, Ruiz, Yannis Kotidis, Steffen Lamparter, Theofilos Mailis, Christian Neuenstadt, Ozgur L. Ozcep, Christoph Pinkel, Christoforos Svingos, Dmitriy Zheleznyakov, Ian Horrocks, Yannis E. Ioannidis and Ralf Moller (2016) Ontology-Based Integration of Streaming and Static Relational Data with Optique,*Proc. of International Conference on Management Data (SIGMOD)*, pp.2109–2112, https://doi.org/10.1145/2882903.2899385.

[76] Paul Groth, Antonis Loizou, Alasdair J.G. Gray, Carole Goble, Lee Harland, Steve Pettifer (2014) API-centric Linked Data integration: The Open PHACTS Discovery Platform case study,*Web Semantics: Science, Services and Agents on the World Wide Web, Volume 29*, pp. 12-18.

[77] Gomez-Cabrero, David and Abugessaisa, Imad and Maier, Dieter and Teschendorff, Andrew and Merkenschlager, Matthias and Gisel, Andreas and Ballestar, Esteban and Bongcam-Rudloff, Erik and Conesa, Ana and Tegner, Jesper (2014) Data integration in the era of omics: current and future challenges, *BMC Systems Biology*, https://doi.org/10.1186/1752-0509-8-S2-I1.

[78] Craig A. Knoblock, Pedro Szekely (2015) Exploiting Semantics for Big Data Integration, *Association for the Advancement of Artificial Intelligence*.

[79] Lu Tang, Peter X.K. Song (2016) Fused Lasso Approach in Regression Coefficients Clustering Learning Parameter Heterogeneity in Data Integration, *Journal of Machine Learning Research*.

[80] Son T. Mai, Xiao He, Jing Feng, Claudia Plant, Christian Bohm (2016) Anytime density-based clustering of complex data, *Knowledge and Information Systems*, pp 319-355.

[81] Diane J. Janvrin, Robyn L. Raschke, William N. Dilla (2014) Making sense of complex data using interactive data visualization,*Journal of Accounting Education, Volume 32, Issue 4*, pp 31-48, https://doi.org/10.1016/j.jaccedu.2014.09.003.

[82] N. Tsagkarakis, P. P. Markopoulos, G. Sklivanitis and D. A. Pados (2018) L1-Norm Principal-Component Analysis of Complex Data, *IEEE Transactions on Signal Processing, vol. 66, no. 12, 15 June15*, pp. 3256-3267, https://doi.org/10.1109/TSP.2018.2821641.

[83] Vladimir Berikov (2014) Weighted ensemble of algorithms for complex data clustering,*Pattern Recognition Letters, Volume 38*, pp 99-106, https://doi.org/10.1016/j.patrec.2013.11.012.

[84] Zuguang Gu Roland Eils Matthias Schlesner (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data,*Bioinformatics, Volume 32, Issue 18*, pp 2847-2849, https://doi.org/10.1093/bioinformatics/btw313.

[85] Thierry Denoeux (2014) Likelihood-based belief function: justification and some extensions to low-quality data ,*International Journal of Approximate Reasoning*, pp.1535-1547, https://doi.org/10.1016/j.ijar.2013.06.007.

[86] Paul W. Murray, Bruno Agard, Marco A. Barajas (2017) Market segmentation through data mining: A method to extract behaviors from a noisy data set,*Computers and Industrial Engineering, Volume 109*, pp 233-252, https://doi.org/10.1016/j.cie.2017.04.017.

[87] Carlos J. Mantas, Joaquin Abellan : Credal-C4.5 (2014) Decision tree based on imprecise probabilities to classify noisy data,*Expert Systems with Applications, Volume 41, Issue 10*, pp.4625-4637, https://doi.org/10.1016/j.eswa.2014.01.017.

[88] B. Frenay and M. Verleysen (2016) Reinforced Extreme Learning Machines for Fast Robust Regression in the Presence of Outliers, *IEEE Transactions on Cybernetics, vol. 46, no. 12*, pp. 3351-3363, https://doi.org/10.1109/TCYB.2015.2504404.

[89] M. Kang, M. R. Islam, J. Kim, J. Kim and M. Pecht (2016) A Hybrid Feature Selection Scheme for Reducing Diagnostic Performance Deterioration Caused by Outliers in Data-Driven Diagnostics,*IEEE Transactions on Industrial Electronics, vol. 63, no. 5*, pp.3299-3310, https://doi.org/10.1109/TIE.2016.2527623.

[90] J. Chen et al (2017) A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment,*IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 4*, pp. 919-933, https://doi.org/10.1109/TPDS.2016.2603511.

[91] I. Chung et al (2017) Parallel Deep Neural Network Training for Big Data on Blue Gene/Q, *IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 6*, pp. 1703-1714, https://doi.org/10.1109/TPDS.2016.2626289.

[92] Jin Qian, Ping Lv, Xiaodong Yue, Caihui Liu, Zhengjun Jing (2015) Hierarchical attribute reduction algorithms for big data using MapReduce,*Knowledge-Based Systems, Volume 73*, pp. 18-31, https://doi.org/10.1016/j.knosys.2014.09.001.

[93] Sergio Ramirez,Gallego,Iago Lastra, David Martinez,Rego, Veronica Bolon,Canedo, Jose Manuel Benitez,Francisco Herrera, Amparo Alonso,Betanzos (2016) Fast,mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High Dimensional Big Data,*International Journal of Intelligent Systems*, https://doi.org/10.1002/int.21833.

[94] S. Wang, L. L. Minku and X. Yao (2015) Resampling-Based Ensemble Methods for Online Class Imbalance Learning, *IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 5*, pp. 1356-1368, https://doi.org/10.1109/TKDE.2014.2345380.

[95] K. E. Bennin, J. Keung, P. Phannachitta, A. Monden and S. Mensah (2018) MAHAKIL- Diversity Based Oversampling Approach to Alleviate the Class Imbalance Issue in Software Defect Prediction, *IEEE Transactions on Software Engineering, vol. 44, no. 6*, pp. 534-550, https://doi.org/10.1109/TSE.2017.2731766.

[96] Loris Nanni, Carlo Fantozzi, Nicola Lazzarini (2015) Coupling different methods for overcoming the class imbalance problem,*Neurocomputing, Volume 158*, pp. 48-61, https://doi.org/10.1016/j.neucom.2015.01.068.

[97] Zhen Liu, Deyu Tang, Yongming Cai, Ruoyu Wang, Fuhua Chen (2017) A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data, *Neurocomputing, Volume 266*, pp. 641-650, https://doi.org/10.1016/j.neucom.2017.05.066.

[98] Tavish Srivastava (2015) Perfect way to build a Predictive Model, Analytics Vidhya, https://www.analyticsvidhya.com.

[99] Jeremy Howick, Paul Glasziou, Jeffrey K. Aronson (2013) Problems with using mechanisms to solve the problem of extrapolation, *Theoretical Medicine and Bioethics, Volume 34, Issue 4*, pp 275–291, https://doi.org/10.1007/s11017-013-9266-0.

[100] Manuel Martin-Flores, Monique D. Parr, Luis Campoy, Robin D. Gleed (2012) The sensitivity of sheep to vecuronium: an example of the limitations of extrapolation, *Canadian Journal of Anesthesia, Volume 59, Issue 7*, pp 722–723, https://doi.org/10.1007/s12630-012-9707-7.

[101] James R. Miller, Monica G. Turner, Erica A. H. Smithwick, C. Lisa Dent, Emily H. Stanley (2004) Spatial Extrapolation: The Science of Predicting Ecological Patterns and Processes, *BioScience, Volume 54, Issue 4*,Pages 310–320, https://doi.org/10.1641/0006-3568(2004)054[0310:SETSOP]2.0.CO;2.

[102] Forbes, V. E., Calow, P. and Sibly, R. M. (2009) The extrapolation problem and how population modeling can help, *Environmental Toxicology and Chemistry*, pp: 1987-1994, https://doi.org/10.1897/08-029.1.

[103] Stack Exchange (2016) What is wrong with Extrapolation, *StackExchange*, https://stats.stackexchange.com/questions.

[104] Peter Flom (2018) The disadvantages of linear regression, *Sciencing*, https://sciencing.com/disadvantages-linear-regression-8562780.html.

[105] Benjamin A. Goldstein, Ann Marie Navar, Rickey E. Carter (2017) Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges, *European Heart Journal, Volume 38, Issue 23*, Pages 1805–1814.

[106] Harvey J Motulsky and Ronald E Brown (2006) Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate, *BMC Bioinformatics*.

[107] Johansen, S., and Nielsen, B (2016) Asymptotic Theory of Outlier Detection Algorithms for Linear Time Series Regression Models. *Scand J Statist*, 43: 321– 348, https://doi.org/10.1111/sjos.12174.

[108] Minitab Blog Editor (2013) Enough Is Enough! Handling Multicollinearity in Regression Analysis, *The Minitab Blog*, https://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in\-regression-analysis.

[109] Vatcheva KP, Lee M, McCormick JB, Rahbar MH (2016) Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale)*,6(2):227.

[110] Jake Lever, Martin Krzywinski and Naomi Altman (2016) Logistic regression, *Nature Methods*, pages 541–542, https://doi.org/10.1038/nmeth.3904.

[111] Carina Mood (2010) Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It, *European Sociological Review*, Volume 26, Issue 1, Pages 67–82, https://doi.org/10.1093/esr/jcp006.

[112] Muchlinski, D., Siroky, D., He, J., and Kocher, M. (2016) Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, 24(1), 87-103, https://doi.org/10.1093/pan/mpv024.

[113] Mirjam J. Knol, Saskia Le Cessie, Ale Algra, Jan P. Vandenbroucke, Rolf H.H. Groenwold (2012) Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression, *CMAJ*, 184 (8) 895-899, https://doi.org/10.1503/cmaj.101715.

[114] Pijush Samui (2013) Multivariate Adaptive Regression Spline (Mars) for Prediction of Elastic Modulus of Jointed Rock Mass, *Geotechnical and Geological Engineering*, Volume 31, Issue 1, pp 249–253, https://doi.org/10.1007/s10706-012-9584-4.

[115] Pijush Samui, Sarat Das, Dookie Kim (2011) Uplift capacity of suction caisson in clay using multivariate adaptive regression spline,*Ocean Engineering, Volume 38, Issues 17–18*, Pages 2123-2127, https://doi.org/10.1016/j.oceaneng.2011.09.036.

[116] Kuhn Max; Johnson Kjell (2013) MARS regression, *Applied Predictive Modeling*, New York, NY: Springer New York.

[117] Statsoft (2019) Multivariate Adaptive Regression Splines (MARSplines), *Statsoft.com*, http://www.statsoft.com/Textbook/Multivariate-Adaptive-Regression-Splines.

[118] Pijush Samui,Pradeep Kurup (2012) Multivariate Adaptive Regression Spline and Least Square Support Vector Machine for Prediction of Undrained Shear Strength of Clay, *International Journal of Applied Metaheuristic Computing, 3(2)*, https://doi.org/10.4018/jamc.2012040103.

[119] Ariadna Montiel, Ruth Lazkoz, Irene Sendra, Celia Escamilla-Rivera, and Vincenzo Salzano (2014) Nonparametric reconstruction of the cosmic expansion with local regression smoothing and simulation extrapolation, *Physical Review D, 89, 043007*, https://doi.org/10.1103/PhysRevD.89.043007.

[120] Felix Biscarri, Inigo Monedero, Antonio Garcia, Juan Ignacio Guerrero, Carlos Leon (2017) Electricity clustering framework for automatic classification of customer loads, *Expert Systems with Applications, Volume 86*, Pages 54-63, https://doi.org/10.1016/j.eswa.2017.05.049.

[121] NIST/SEMATECH (2012) e-Handbook of Statistical Methods, *NIST/SEMATECH*, http://www.itl.nist.gov/div898/handbook.

[122] Kyoosik Kim (2019) Ridge Regression for Better Usage, *Towards data science*, https://towardsdatascience.com/ridge-regression-for-better-usage.

[123] C.B. Garcia, J. Garcia, M.M. Lopez Martín and R. Salmeron (2015) Collinearity: revisiting the variance inflation factor in ridge regression, *Journal of Applied Statistics, Volume 42 - Issue 3*, Pages 648-661, https://doi.org/10.1080/02664763.2014.980789.

[124] NCSS Statistical Software: Ridge regression, *NCSS*, https://ncss-wpengine.netdna-ssl.com.

[125] Patrick Breheny: Penalized regression methods, *University of Kentucky*, https://web.as.uky.edu/statistics/users/pbreheny.

[126] B M Golam Kibria, Shipra Banik (2016) Some Ridge Regression Estimators and Their Performances, *Journal of Modern Applied Statistical Methods, Volume 15, Issue 1 Article 12*.

[127] Prashant Gupta (2017) Regularization in Machine Learning, *Towards data science*, https://towardsdatascience.com/regularization-in-machine-learning.

[128] Gene H. Golub, Michael Heath and Grace Wahba (2012) Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter, *Technometrics, 21:2*, 215-223, https://doi.org/10.1080/00401706.1979.10489751.

[129] Charles K. Fisher, Austin Huang, and Collin M. Stultz (2010) Modeling Intrinsically Disordered Proteins with Bayesian Statistics, *Journal of the American Chemical Society, 132 (42)*, 14919-14927, https://doi.org/10.1021/ja105832g.

[130] Andre E. Punt and Ray Hilborn (2001) Strengths and weaknesses of Bayesian Approach, *Computerized Information Series*, Food and Organization of the United Nations.

[131] Hoffrage Ulrich, Krauss Stefan, Martignon Laura, Gigerenzer Gerd (2015) Natural frequencies improve Bayesian reasoning in simple and complex inference tasks, *Frontiers in Psychology, Volume 6*, pp.1473, https://doi.org/10.3389/fpsyg.2015.01473.

[132] Tina R. Patil, Mrs. S. S. Sherekar, Sant Gadgebaba (2013) Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, *International Journal Of Computer Science And Applications Vol. 6, No.2*.

[133] Narayanan V., Arora I., Bhatia A. (2013) Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model, *Intelligent Data Engineering and Automated Learning – IDEAL 2013. IDEAL*, https://doi.org/10.1007/978-3-642-41278-3_24.

[134] S.L. Ting, W.H. Ip, Albert H.C. Tsang (2011) Is Naïve Bayes a Good Classifier for Document Classification?, *International Journal of Software Engineering and Its Applications Vol. 5, No. 3*.

[135] J. Zhang, C. Chen, Y. Xiang, W. Zhou and Y. Xiang (2013) Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions, *IEEE Transactions on Information Forensics and Security, vol. 8, no. 1*, pp. 5-15, https://doi.org/10.1109/TIFS.2012.2223675.

[136] Nayyar A. Zaidi, Jesus Cerquides, Mark J. Carman, Geoffrey I.Webb (2013) Alleviating Naive Bayes Attribute Independence Assumption by AttributeWeighting, *Journal of Machine Learning Research 14* , 1947-1988.

[137] Liangxiao Jiang, Chaoqun Li, Shasha Wang, Lungan Zhang (2016) Deep feature weighting for naive Bayes and its application to text classification,*Engineering Applications of Artificial Intelligence, Volume 52*, Pages 26-39, https://doi.org/10.1016/j.engappai.2016.02.002.

[138] Victor Roman (2019) Naive Bayes Algorithm: Intuition and Implementation in a Spam Detector, *Towards data science*, https://towardsdatascience.com/naive-bayes-intuition-and-implementation.

[139] Liangxiao Jiang, Zhihua Cai, Dianhong Wang, Harry Zhang (2012) Improving Tree augmented Naive Bayes for class probability estimation,*Knowledge-Based Systems, Volume 26*, Pages 239-245, https://doi.org/10.1016/j.knosys.2011.08.010.

[140] Zhun Yu, Fariborz Haghighat, Benjamin C.M. Fung, Hiroshi Yoshino (2010) A decision tree method for building energy demand modeling, *Energy and Buildings, Volume 42, Issue 10*, Pages 1637-1646, https://doi.org/10.1016/j.enbuild.2010.04.006.

[141] V.F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, J.P. Rigol-Sanchez (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification, *ISPRS Journal of Photogrammetry and Remote Sensing, Volume 67*, Pages 93-104, https://doi.org/10.1016/j.isprsjprs.2011.11.002.

[142] Mahyat Shafapour Tehrany, Biswajeet Pradhan, Mustafa Neamah Jebur (2013) Spatial prediction of

flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS, *Journal of Hydrology, Volume 504*, Pages 69-79, https://doi.org/10.1016/j.jhydrol.2013.09.034.

[143] Jung Hwan Cho, Pradeep U. Kurup (2011) Decision tree approach for classification and dimensionality reduction of electronic nose data, *Sensors and Actuators B: Chemical, Volume 160, Issue 1*, Pages 542-548, https://doi.org/10.1016/j.snb.2011.08.027.

[144] Song YY, Lu Y (2015) Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry. 27(2)*, pp. 130 – 135.

[145] Bartosz Krawczyk, Michał Woźniak, Gerald Schaefer (2014) Cost-sensitive decision tree ensembles for effective imbalanced classification, *Applied Soft Computing, Volume 14, Part C*, Pages 554-562, https://doi.org/10.1016/j.asoc.2013.08.014.

[146] Tao Wang, Zhenxing Qin, Zhi Jin, Shichao Zhang (2010) Handling over-fitting in test cost-sensitive decision tree learning by feature selection, smoothing and pruning, *Journal of Systems and Software, Volume 83, Issue 7*, Pages 1137-1147, https://doi.org/10.1016/j.jss.2010.01.002.

[147] Rutvija Pandya, Jayati Pandya (2015) C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning, *International Journal of Computer Applications Volume 117 – No. 16*, https://doi.org/10.5120/20639-3318.

[148] Shengyi Jiang, Guansong Pang, Meiling Wu, Limin Kuang (2012) An improved K-nearest-neighbor algorithm for text categorization, *Expert Systems with Applications, Volume 39, Issue 1*, Pages 1503-1509, https://doi.org/10.1016/j.eswa.2011.08.040.

[149] Sadegh Bafandeh Imandoust And Mohammad Bolandraftar (2013) Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background, *S B Imandoust et al. Int. Journal of Engineering Research and Applications, Vol. 3, Issue 5*, pp.605-610.

[150] X. Liang, X. Gou and Y. Liu (2012) Fingerprint-based location positoning using improved KNN, *2012 3rd IEEE International Conference on Network Infrastructure and Digital Content, Beijing*, pp. 57-61.

[151] A. Thommandram, J. M. Eklund and C. McGregor (2013) Detection of apnoea from respiratory time series data using clinically recognizable features and kNN classification, *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka*, pp. 5013-5016, https://doi.org/10.1109/EMBC.2013.6610674.

[152] Carmelo Cassisi, Alfredo Ferro, Rosalba Giugno, Giuseppe Pigola, Alfredo Pulvirenti (2013) Enhancing density-based clustering: Parameter reduction and outlier detection, *Information Systems, Volume 38, Issue 3*, Pages 317-330, https://doi.org/10.1016/j.is.2012.09.001.

[153] O. Kursun (2010) Spectral Clustering with Reverse Soft K-Nearest Neighbor Density Estimation, *The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona*, pp. 1-8.

[154] aporras (2018) 10 Reasons for loving Nearest Neighbors algorithm, *QuantDare*, https://quantdare.com/10-reasons\-for-nearest-neighbors-algorithm.

[155] Y. Mitani, Y. Hamamoto (2006) A local mean-based nonparametric classifier, *Pattern Recognition Letters, Volume 27, Issue 10*, Pages 1151-1159, https://doi.org/10.1016/j.patrec.2005.12.016.

[156] Gustavo E.A.P.A. Batista, Diego Furtado Silva (2009) How k-Nearest Neighbor Parameters Affect its Performance?, *Argentine symposium on artificial intelligence*.

[157] Shichao Zhang, Debo Cheng, Zhenyun Deng, Ming Zong, Xuelian Deng (2018) A novel kNN algorithm with data-driven k parameter computation, *Pattern Recognition Letters, Volume 109*, Pages 44-54, https://doi.org/10.1016/j.patrec.2017.09.036.

[158] Zhang, Shichao and Li, Xuelong and Zong, Ming and Zhu, Xiaofeng and Cheng, Debo (2017) Learning K for kNN Classification, *ACM Trans. Intell. Syst. Technol., Volume 8*, pp. 43:1–43:19, https://doi.org/10.1145/2990508.

[159] Abdulhamit Subasi (2013) Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders, *Computers in Biology and Medicine, Volume 43, Issue 5*, Pages 576-586, https://doi.org/10.1016/j.compbiomed.2013.01.020.

[160] Fangjun Kuang, Weihong Xu, Siyang Zhang (2014) A novel hybrid KPCA and SVM with GA model for intrusion detection,*Applied Soft Computing, Volume 18*, Pages 178-184, https://doi.org/10.1016/j.asoc.2014.01.028.

[161] Roman M. Balabin, Ekaterina I. Lomakina (2011) Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data , *Analyst, issue:8*.

[162] S. Han, Cao Qubo and Han Meng (2012) Parameter selection in SVM with RBF kernel function, *World Automation Congress 2012, Puerto Vallarta, Mexico*, pp. 1-4.

[163] Weijun li, Zhenyu Liu (2011) A method of SVM with Normalization in Intrusion Detection, *Procedia Environmental Sciences, Volume 11, Part A*, Pages 256-262, https://doi.org/10.1016/j.proenv.2011.12.040.

[164] Adel Bolbol, Tao Cheng, Ioannis Tsapakis, James Haworth (2012) Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification, *Computers, Environment and Urban Systems, Volume 36, Issue 6*, Pages 526-537, https://doi.org/10.1016/j.compenvurbsys.2012.06.001.

[165] A. Bhardwaj, A. Gupta, P. Jain, A. Rani and J. Yadav (2015) Classification of human emotions from EEG signals using SVM and LDA Classifiers, *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN), Noida*, pp. 180-185, https://doi.org/10.1109/SPIN.2015.7095376.

[166] De Giorgi, M.G., Campilongo, S., Ficarella, A., Congedo, P.M (2014) Comparison Between Wind Power Prediction Models Based on Wavelet Decomposition with Least-Squares Support Vector Machine (LS-SVM) and Artificial Neural Network (ANN) *Energies, 7*, 5251-5272, https://doi.org/10.3390/en7085251.

[167] M. Emre Celebi, Hassan A. Kingravi, Patricio A. Vela (2013) A comparative study of efficient initialization methods for the k-means clustering algorithm, *Expert Systems with Applications, Volume 40, Issue 1* , Pages 200-210, https://doi.org/10.1016/j.eswa.2012.07.021.

[168] Trupti M. Kodinariya, Prashant R. Makwana (2013) Review on determining number of Cluster in K-Means Clustering, *International Journal of Advance Research in Computer Science and Management Studies, Volume 1, Issue 6*.

[169] S. Na, L. Xumin and G. Yong (2010) Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm, *Third International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan*, pp. 63-67.

[170] Soumi Ghosh, Sanjay Kumar Dubey (2013) Comparative analysis of k-means and fuzzy c-means algorithms, *International Journal of Advanced Computer Science and Applications, Vol. 4, No.4*, https://doi.org/10.14569/IJACSA.2013.040406.

[171] Taher Niknam, Babak Amiri (2010) An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis,*Applied Soft Computing, Volume 10, Issue 1* , pp. 183-197, https://doi.org/10.1016/j.asoc.2009.07.001.

[172] You Li, Kaiyong Zhao, Xiaowen Chu, Jiming Liu (2013) Speeding up k-Means algorithm by GPUs,*Journal of Computer and System Sciences, Volume 79, Issue 2*, pp. 216-229, https://doi.org/10.1016/j.jcss.2012.05.004.

[173] Steinley, D., and Brusco, M. J (2011) Choosing the number of clusters in K-means clustering, *Psychological Methods, 16(3)*, 285-297, https://doi.org/10.1037/a0023346.

[174] Renato Cordeiro de Amorim, Boris Mirkin (2012) Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering, *Pattern Recognition, Volume 45, Issue 3* , pp. 1061-1075, https://doi.org/10.1016/j.patcog.2011.08.012.

[175] Carlos Ordonez, Edward Omiecinski (2004) Efficient Disk-Based K-Means Clustering for Relational Databases, *IEEE Transactions on Knowledge and Data Engineering, vol. 16*, pp. 909-921, https://doi.org/10.1109/TKDE.2004.25.

[176] Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, Andy Song (2015) Efficient agglomerative hierarchical clustering, *Expert Systems with Applications, Volume 42, Issue 5*, pp 2785-2797, https://doi.org/10.1016/j.eswa.2014.09.054.

[177] Dongkuan Xu, Yingjie Tian (2015) A Comprehensive Survey of Clustering Algorithms, *Annals of Data Science, Volume 2, Issue 2*, pp 165–193, https://doi.org/10.1007/s40745-015-0040-1.

[178] Shafiq Alam, Gillian Dobbie, Yun Sing Koh, Patricia Riddle, Saeed Ur Rehman (2014) Research on particle swarm optimization based clustering: A systematic review of literature and techniques, *Swarm and Evolutionary Computation, Volume 17*, Pages 1-13, https://doi.org/10.1016/j.swevo.2014.02.001.

[179] T. Nguyen and C. Kwoh (2015) Efficient agglomerative hierarchical clustering for biological sequence analysis, *TENCON 2015 - 2015 IEEE Region 10 Conference, Macao*, pp. 1-3.

[180] Guilherme Andrade, Gabriel Ramos, Daniel Madeira, Rafael Sachetto, Renato Ferreira, Leonardo Rocha (2013) G-DBSCAN: A GPU Accelerated Algorithm for Density-based Clustering, *Procedia Computer Science, Volume 18*, Pages 369-378, https://doi.org/10.1016/j.procs.2013.05.200.

[181] Younghoon Kim, Kyuseok Shim, Min-Soeng Kim, June Sup Lee (2014) DBCURE-MR: An efficient density-based clustering algorithm for large data using MapReduce, *Information Systems, Volume 42*, Pages 15-35, https://doi.org/10.1016/j.is.2013.11.002.

[182] Carmelo Cassisi, Alfredo Ferro, Rosalba Giugno, Giuseppe Pigola, Alfredo Pulvirenti (2013) Enhancing density-based clustering: Parameter reduction and outlier detection, *Information Systems, Volume 38, Issue 3*, Pages 317-330, https://doi.org/10.1016/j.is.2012.09.001.

[183] Sunita Jahirabadkar, Parag Kulkarni (2014) Algorithm to determine ϵ-distance parameter in density based clustering, *Expert Systems with Applications, Volume 41, Issue 6*, Pages 2939-2946, https://doi.org/10.1016/j.eswa.2013.10.025.

[184] A Amini, TY Wah (2011) Density micro-clustering algorithms on data streams: A review, *Proceedings of the International Multiconference of Engineers and Computer Scientists*.

[185] Dongwon Lee, Sung-Hyuk Park, Songchun Moon (2013) Utility-based association rule mining: A marketing solution for cross-selling, *Expert Systems with Applications, Volume 40, Issue 7*, Pages 2715-2725, https://doi.org/10.1016/j.eswa.2012.11.021.

[186] Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen (2013) Association rule mining to detect factors which contribute to heart disease in males and females, *Expert Systems with Applications, Volume 40, Issue 4*, Pages 1086-1093, https://doi.org/10.1016/j.eswa.2012.08.028.

[187] R.J. Kuo, C.M. Chao, Y.T. Chiu (2011) Application of particle swarm optimization to association rule mining, *Applied Soft Computing, Volume 11, Issue 1*, Pages 326-336, https://doi.org/10.1016/j.asoc.2009.11.023.

[188] Zhang M., He C. (2010) Survey on Association Rules Mining Algorithms , *Advancing Computing, Communication, Control and Management. Lecture Notes in Electrical Engineering, vol 56. Springer, Berlin, Heidelberg.*, pp 111-118, https://doi.org/10.1007/978-3-642-05173-9_15.

[189] Andrew Tch: The mostly complete chart of Neural Networks, explained, *Towards data science*, https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464.

[190] H. Yu, T. Xie, S. Paszczynski and B. M. Wilamowski (2011) Advantages of Radial Basis Function Networks for Dynamic System Design, *IEEE Transactions on Industrial Electronics, vol. 58, no. 12*, pp. 5438-5450, https://doi.org/10.1109/TIE.2011.2164773.

[191] Mehdi Khashei, Mehdi Bijari (2011) A novel hybridization of artificial neural networks and ARIMA models for time series forecasting, *Applied Soft Computing, Volume 11, Issue 2*, Pages 2664-2675, https://doi.org/10.1016/j.asoc.2010.10.015.

[192] Li-Hua Feng, Jia Lu (2010) The practical research on flood forecasting based on artificial neural networks, *Expert Systems with Applications, Volume 37, Issue 4*, Pages 2974-2977, https://doi.org/10.1016/j.eswa.2009.09.037.

[193] Daniel Westreich, Justin Lessler, Michele Jonsson Funk (2010) Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression, *Journal of Clinical Epidemiology, Volume 63, Issue 8*, Pages 826-833, https://doi.org/10.1016/j.jclinepi.2009.11.020.

[194] Ding, S., Su, C. and Yu (2011) An optimizing BP neural network algorithm based on genetic algorithm*Artificial Intelligence Review , Volume 36, Issue 2*, pp 153–162, https://doi.org/10.1007/s10462-011-9208-z.

[195] Hong-ze Li, Sen Guo, Chun-jie Li, Jing-qi Sun (2013) A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm, *Knowledge-Based Systems, Volume 37*, Pages 378-387, https://doi.org/10.1016/j.knosys.2012.08.015.

[196] SeyedAli Mirjalili, Siti Zaiton Mohd Hashim, Hossein Moradian Sardroudi (2012) Training feedforward neural networks using hybrid particle swarm optimization and gravitational search algorithm, *Applied Mathematics and Computation, Volume 218, Issue 22*, Pages 11125-11137, https://doi.org/10.1016/j.amc.2012.04.069.

[197] N Srivastava, G Hinton, A Krizhevsky, I Sutskever, R Salakhutdinov (2014) Dropout: a simple way to prevent neural networks from overfitting, *Journal of machine learning research*.

[198] Jonathan L. Ticknor (2013) A Bayesian regularized artificial neural network for stock market forecasting, *Expert Systems with Applications, Volume 40, Issue 14*, Pages 5501-5506, https://doi.org/10.1016/j.eswa.2013.04.013.

[199] Jurgen Schmidhuber (2015) Deep learning in neural networks: An overview, *Neural Networks, Volume 61*, Pages 85-117, https://doi.org/10.1016/j.neunet.2014.09.003.

[200] Chao Shang, Fan Yang, Dexian Huang, Wenxiang Lyu (2014) Data-driven soft sensor development based on deep learning technique, *Journal of Process Control, Volume 24, Issue 3*, Pages 223-233, https://doi.org/10.1016/j.jprocont.2014.01.012.

[201] Jie Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen and Chung-Ming Chen (2016)

Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans, *Scientific Reports volume 6*, https://doi.org/10.1038/srep24454.

[202] Ravid Shwartz-Ziv, Naftali Tishby (2017) Opening the Black Box of Deep Neural Networks via Information, *Machine learning*, https://arxiv.org/abs/1703.00810.

[203] Tanu Arya (2018) Drawbacks of Deep Learning, *Stanford Management Science and Engineering*, https://mse238blog.stanford.edu.

[204] P.K. Anooj (2012) Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules, *Journal of King Saud University - Computer and Information Sciences, Volume 24, Issue 1*, Pages 27-40, https://doi.org/10.1016/j.jksuci.2011.09.002.

[205] Jose M. Alonso, Luis Magdalena (2011) Special issue on interpretable fuzzy systems, *Information Sciences, Volume 181, Issue 20*, Pages 4331-4339, https://doi.org/10.1016/j.ins.2011.07.001.

[206] Salma Elhag, Alberto Fernández, Abdullah Bawakid, Saleh Alshomrani, Francisco Herrera (2015) On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on Intrusion Detection Systems,*Expert Systems with Applications, Volume 42, Issue 1*, Pages 193-202, https://doi.org/10.1016/j.eswa.2014.08.002.

[207] V. P. G. Jimenez, Y. Jabrane, A. G. Armada, B. Ait Es Said and A. Ait Ouahman (2011) High Power Amplifier Pre-Distorter Based on Neural-Fuzzy Systems for OFDM Signals, *IEEE Transactions on Broadcasting, vol. 57, no. 1*, pp. 149-158, https://doi.org/10.1109/TBC.2010.2088331.

[208] TE Alhanafy, F Zaghlool, A Saad, ED Moustafa (2010) Neuro-Fuzzy modeling scheme for the prediction of air pollution, *Journal of American Science, 6(12)*.

[209] Ilija Svalina, Vjekoslav Galzina, Roberto Lujic, Goran Simunovic (2013) An adaptive network-based fuzzy inference system (ANFIS) for the forecasting: The case of close price indices,*Expert Systems with Applications, Volume 40, Issue 15*, Pages 6055-6063, https://doi.org/10.1016/j.eswa.2013.05.029.

[210] B. Dennis, S. Muthukrishnan (2014) AGFS: Adaptive Genetic Fuzzy System for medical data classification, *Applied Soft Computing, Volume 25*, Pages 242-252, https://doi.org/10.1016/j.asoc.2014.09.032.

[211] P Amudha, S Karthik, S Sivakumari (2013) Classification techniques for intrusion detection an overview , *International Journal of Computer applications, Volume 76, No.16*, https://doi.org/10.5120/13334-0928.

[212] T. M. Khoshgoftaar, J. Van Hulse and A. Napolitano (2011) Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data, *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 41, no. 3*, pp. 552-568, https://doi.org/10.1109/TSMCA.2010.2084081.

[213] Syarif I., Zaluska E., Prugel-Bennett A., Wills G (2012) Application of Bagging, Boosting and Stacking to Intrusion Detection, *International workshop on Machine Learning and Data Mining in Pattern Recognition. MLDM . Lecture Notes in Computer Science, vol 7376. Springer, Berlin, Heidelberg*, https://doi.org/10.1007/978-3-642-31537-4_46.

[214] GeeksforGreeks: Comparison b/w Bagging and Boosting in Data Mining, *https://www.geeksforgeeks.org/comparison-b-w-bagging-and-boosting-data-mining*.

[215] M. Paz Sesmero, Agapito I. Ledezma, Araceli Sanchis (2015) Generating ensembles of heterogeneous classifiers using Stacked Generalization, *Wires data mining and knowledge discovery*.

[216] Harshdeep Singh (2018) Understanding Gradient Boosting Machines, *https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab*.

[217] Ferreira A.J., Figueiredo M.A.T. : Boosting Algorithms: A Review of Methods, Theory, and Applications, *Ensemble Machine Learning. Springer, Boston, MA*, pp 35-85, https://doi.org/10.1007/978-1-4419-9326-7_2.

[218] Wilson, Andrew G and Gilboa, Elad and Nehorai, Arye and Cunningham, John P (2014) Fast Kernel Learning for Multidimensional Pattern Extrapolation, *Advances in Neural Information Processing Systems 27*, pp.3626- 3634.

[219] Wang S () CyberGIS and spatial data science, *GeoJournal, Volume 81, Issue 6*, pp.965–968, https://doi.org/10.1007/s10708-016-9740-0.

[220] William Q. Meeker and Yili Hong (2014) Reliability Meets Big Data: Opportunities and Challenges, *Quality Engineering, 26:1*, 102-116, https://doi.org/10.1080/08982112.2014.846119.

[221] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais and Prabhat (2019) Deep learning and process understanding for data-driven Earth system science, *Nature, volume 566*, pages195–204, https://doi.org/10.1038/s41586-019-0912-1.

[222] Peter V. Coveney, Edward R. Dougherty and Roger R. Highfield (2016) Big data need big theory too, *Philosophical transactions of the Royal Society A Mathematical, Physical and Engineering sciences*.

[223] Xiang Liu, Ziyang Tang, Huyunting Huang, Tonglin Zhang and Baijian Yang (2019) Multiple Learning for Regression in big data, *CoRR*.

[224] Ping Ma, Xiaoxiao Sun (2014) Leveraging for big data regression, *Wires Computational Statistics*.

[225] S Jun, SJ Lee, JB Ryu (2015) A divided regression analysis for big data, *International Journal of software engineering and its applications*.

[226] HaiYing Wang, Min Yang and John Stufken (2019) Information-Based Optimal Subdata Selection for Big Data Linear Regression, *Journal of the American Statistical Association, volume 114, number 525*, pp. 393-405, https://doi.org/10.1080/01621459.2017.1408468.

[227] Yang, Hang and Fong, Simon (2012) Incrementally Optimized Decision Tree for Noisy Big Data, *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*,pp.36-44.

[228] W Dai, W Ji (2014) A mapreduce implementation of C4. 5 decision tree algorithm, *International journal of database theory and application, Vol 7, No. 1*, pp.49-60, https://doi.org/10.14257/ijdta.2014.7.1.05.

[229] J. Chen et al. (2017) A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment, *IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 4*, pp. 919-933, https://doi.org/10.1109/TPDS.2016.2603511.

[230] Ke, Guolin and Meng, Qi and Finley, Thomas and Wang, Taifeng and Chen, Wei and Ma, Weidong and Ye, Qiwei and Liu, Tie-Yan (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Advances in Neural Information Processing Systems 30*, pp. 3146-3154.

[231] Zhenyun Deng, Xiaoshu Zhu, Debo Cheng, Ming Zong, Shichao Zhang (2016) Efficient kNN classification algorithm for big data, *Neurocomputing, Volume 195*, Pages 143-148, https://doi.org/10.1016/j.neucom.2015.08.112.

[232] Jesus Maillo, Sergio Ramírez, Isaac Triguero, Francisco Herrera (2017) kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data,*Knowledge-Based Systems, Volume 117*, Pages 3-15, https://doi.org/10.1016/j.knosys.2016.06.012.

[233] J. Maillo, I. Triguero and F. Herrera (2015) A MapReduce-Based k-Nearest Neighbor Approach for Big Data Classification, *IEEE Trustcom/BigDataSE/ISPA, Helsinki*, pp. 167-172.

[234] X Yan, Z Wang, D Zeng, C Hu and H Yao (2014) Design and analysis of parallel MapReduce based KNN-join algorithm for big data classification, *TELKOMNIKA Indonesian Journal of ELectrical Engineering, Vol 12, No 11* , pp.7927-7934, https://doi.org/10.11591/telkomnika.v12i11.6357.

[235] G. Song, J. Rochas, L. E. Beze, F. Huet and F. Magoules (2016) K Nearest Neighbour Joins for Big Data on MapReduce: A Theoretical and Experimental Analysis, *IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 9*, pp. 2376-2392, https://doi.org/10.1109/TKDE.2016.2562627.

[236] Katkar, V. D., and Kulkarni, S. V. (2013) A novel parallel implementation of Naive Bayesian classifier for Big Data, *International Conference on Green Computing, Communication and Conservation of Energy (ICGCE)*

[237] B. Chandra, Manish Gupta (2011) Robust approach for estimating probabilities in Naïve–Bayes Classifier for gene expression data, *Expert Systems with Applications,Volume 38, Issue 3*,Pages 1293-1298, https://doi.org/10.1016/j.eswa.2010.06.076.

[238] Zhou, L., Pan, S., Wang, J., and Vasilakos, A. V. (2017) Machine learning on big data: Opportunities and challenges, *Neurocomputing, 237*, pp.350–361, https://doi.org/10.1016/j.neucom.2017.01.026.

[239] Sergio Ramirez Gallego ,Salvador Garcia, Hector Mourino-Talin , David Martinez-Rego , Veronica Bolon-Canedo ,Amparo Alonso-Betanzos ,Jose Manuel Benitez ,Francisco Herrera (2015) Data discretization: taxonomy and big data challenge, *Advanced Review – Wires Data mining and knowledge discovery*, https://doi.org/10.1002/widm.1173.

[240] Rebentrost, Patrick and Mohseni, Masoud and Lloyd, Seth (2014) Quantum Support Vector Machine for Big Data Classification, *Phys. Rev. Lett. – Volume 113, Issue 13, pp. 130503*, https://doi.org/10.1103/PhysRevLett.113.130503.

[241] Anushree Priyadarshini, SonaliAgarwal (2015) A Map Reduce based Support Vector Machine for Big Data Classification, *International Journal of Database Theory and Application , Vol.8 No.5* ,pp.77-98, https://doi.org/10.14257/ijdta.2015.8.5.07 .

[242] D. Singh, D. Roy and C. K. Mohan : DiP-SVM (2017) Distribution Preserving Kernel Support Vector Machine for Big Data, *IEEE Transactions on Big Data, vol. 3, no. 1*, pp. 79-90, https://doi.org/10.1109/TBDATA.2016.2646700.

[243] Cavallaro, G., Riedel, M., Richerzhagen, M., Benediktsson, J. A., and Plaza, A (2015) On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(10)*, 4634–4646, https://doi.org/10.1109/JSTARS.2015.2458855.

[244] Y. Liu and J. Du (2015) Parameter Optimization of the SVM for Big Data, *8th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou*, pp. 341-344.

[245] Xiao Cai, Feiping Nie, Heng Huang (2013) Multi-View K-Means Clustering on Big Data, *Twenty-Third International Joint Conference on Artificial Intelligence, Web and Knowledge-Based Information Systems*.

[246] Cui, X., Zhu, P., Yang, X., Li, K., and Ji, C. (2014) Optimized big data K-means clustering using MapReduce, *The Journal of Supercomputing, 70(3)*, pp.1249–1259, https://doi.org/10.1007/s11227-014-1225-7.

[247] N. Akthar, M. V. Ahamad and S. Khan (2015) Clustering on Big Data Using Hadoop MapReduce, *International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur*, pp. 789-795.

[248] Rathore, Punit and Kumar, Dheeraj and C. Bezdek, James and Rajasegarar, Sutharshan and Palaniswami, Marimuthu (2018) A Rapid Hybrid Clustering Algorithm for Large Volumes of High Dimensional Data , *IEEE Transactions on Knowledge and Data Engineering* PP. 10.1109.

[249] T. C. Havens, J. C. Bezdek and M. Palaniswami (2013) Scalable single linkage hierarchical clustering for big data, *IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Melbourne, VIC*, pp. 396-401.

[250] Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, Andy Song (2015) Efficient agglomerative hierarchical clustering, *Expert Systems with Applications,Volume 42, Issue 5*,Pages 2785-2797, https://doi.org/10.1016/j.eswa.2014.09.054.

[251] Yunpeng Cai, Yijun Sun, ESPRIT-Tree (2011) Hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time, *Nucleic Acids Research, Volume 39, Issue 14*, Page e95, https://doi.org/10.1093/nar/gkr349.

[252] Embrechts, M and Gatti, Christopher and Linton, Jonathan and Roysam, Badrinath (2013) Hierarchical Clustering for Large Data Sets , *Advances in Intelligent Signal Processing and Data Mining: Theory and Applications*, pp.197-233, https://doi.org/10.1007/978-3-642-28696-4_8.

[253] Yanwei Yu, Jindong Zhao,Xiaodong Wang, Qin Wang , Yonggang Zhang (2015) Cludoop: An Efficient Distributed Density-Based Clustering for Big Data Using Hadoop, *International Journal of Distributed Sensor Networks*.

[254] Amini, A., Wah, T.Y. and Saboohi H. J (2014) On Density-Based Data Streams Clustering Algorithms: A Survey, *Journal of Computer Science and Technology - Volume 29, Issue 1*, pp 116–141, https://doi.org/10.1007/s11390-014-1416-y.

[255] Younghoon Kim, Kyuseok Shim, Min-Soeng Kim, June Sup Lee (2014) DBCURE-MR: An efficient density-based clustering algorithm for large data using MapReduce, *Information Systems, Volume 42*,Pages 15-35, https://doi.org/10.1016/j.is.2013.11.002.

[256] Deng, Z., Hu, Y., Zhu, M. et al (2015) A scalable and fast OPTICS for clustering trajectory big data, *Cluster Computing 18: 549*, https://doi.org/10.1007/s10586-014-0413-9.

[257] Imran Khan, Joshua Z. Huang, Kamen Ivanov (2016) Incremental density-based ensemble clustering over evolving data streams,*Neurocomputing, Volume 191*,Pages 34-43, https://doi.org/10.1016/j.neucom.2016.01.009.

[258] Wang, Yu and Li, Boxun and Luo, Rong and Chen, Yiran and Xu, Ningyi and Yang, Huazhong (2014) Energy efficient neural networks for big data analytics, *Proceedings of the Conference on Design, Automation and Test in Europe*.

[259] Cao J, Cui H, Shi H, Jiao L : Big Data (2016) A Parallel Particle Swarm Optimization Back Propagation Neural Network Algorithm Based on MapReduce,*PLoS ONE 11(6): e0157551*, https://doi.org/10.1371/journal.pone.0157551.

[260] Chiroma, Haruna, Ali Abdullahi, Usman, Abdulhamid, Shafi i, Abdulsalam AlArood, Ala and Gabralla, Lubna and Rana, Nadim and Shuib, Liyana and Hashem, Ibrahim and Dada, Emmanuel and Abubakar, Adamu and Zeki, Akram and Herawan, Tutut (2018) Progress on Artificial Neural Networks for Big Data Analytics: A Survey, *IEEE Access. PP. 1-1*.

[261] Zhang, Y Guo, Q Wang, J : Big data analysis using neural networks 49. 9-18, *10.15961/j.jsuese.2017.01.002*.

[262] Hai Wang, Zeshui Xu, Witold Pedrycz (2017) An overview on the roles of fuzzy set techniques in big data processing: Trends, challenges and opportunities, *Knowledge-Based Systems, Volume 118*, Pages 15-30, https://doi.org/10.1016/j.knosys.2016.11.008.

[263] Alberto Fernandez and Cristobal Jose Carmona and Maria Jose del Jesus and Francisco Herrera (2016) A View on Fuzzy Systems for Big Data: Progress and Opportunities, *International Journal of Computational Intelligence Systems, Volume 9*, pp.69-80, https://doi.org/10.1080/18756891.2016.1180820.

[264] X. Chen and X. Lin (2014) Big Data Deep Learning: Challenges and Perspectives, *IEEE Access, vol. 2*, pp. 514-525, https://doi.org/10.1109/ACCESS.2014.2325029.

[265] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald and Edin Muharemagic (2015) Deep learning applications and challenges in big data analytics, *Journal of Big Data, 2:1*, https://doi.org/10.1186/s40537-014-0007-7.

[266] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, Joel T Dudley (2018) Deep learning for healthcare: review, opportunities and challenges, *Briefings in Bioinformatics, Volume 19, Issue 6*, Pages 1236–1246, https://doi.org/10.1093/bib/bbx044.

[267] Bartosz Krawczyk, Leandro L. Minku, Joao Gama, Jerzy Stefanowski, Michal Wozniak (2017) Ensemble learning for data stream analysis: A survey, *Information Fusion,Volume 37*,Pages 132-156, https://doi.org/10.1016/j.inffus.2017.02.004.

[268] Shan Huang, Botao Wang, Junhao Qiu, Jitao Yao, Guoren Wang, Ge Yu (2016) Parallel ensemble of online sequential extreme learning machine based on MapReduce, *Neurocomputing,Volume 174, Part A*,Pages 352-367, https://doi.org/10.1016/j.neucom.2015.04.105.

[269] Huang X., Ye Y., Zhang H. (2016) Extending Kmeans-Type Algorithms by Integrating Intra-cluster Compactness and Inter-cluster Separation, *Unsupervised Learning Algorithms. Springer*, pp 343-384, https://doi.org/10.1007/978-3-319-24211-8_13.

[270] N. B. Nikhare and P. S. Prasad (2018) A review on inter-cluster and intra-cluster similarity using bisected fuzzy C-mean technique via outward statistical testing, *2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore*, pp. 215-217.

[271] Frederic Godin, Jonas Degrave, Joni Dambre, Wesley De Neve (2018) Dual Rectified Linear Units (DReLUs): A replacement for tanh activation functions in Quasi-Recurrent Neural Networks, *Pattern Recognition Letters, Volume 116*, Pages 8-14, https://doi.org/10.1016/j.patrec.2018.09.006.

[272] Yu Wang (2017) A new concept using LSTM Neural Networks for dynamic system identification, *American Control Conference (ACC), Seattle, WA*, pp. 5324-5329, https://doi.org/10.23919/ACC.2017.7963782.

[273] Igor M. Coelho, Vitor N. Coelho, Eduardo J. da S. Luz, Luiz S. Ochi, Frederico G. Guimaraes, Eyder Rios (2017) A GPU deep learning metaheuristic based model for time series forecasting, Applied Energy, Volume 201, Pages 412-418, https://doi.org/10.1016/j.apenergy.2017.01.003.

[274] Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua (2014) Generative Adversarial Networks (PDF). Proceedings of the International Conference on Neural Information Processing Systems (NIPS). pp. 2672–2680.

[275] Michael A Farnum, Lalit Mohanty, Mathangi Ashok, Paul Konstant, Joseph Ciervo, Victor S Lobanov, Dimitris K Agrafiotis (2019) A dimensional warehouse for integrating operational data from clinical trials, *Database, Volume 2019*, https://doi.org/10.1093/database/baz039.

[276] Khan S.I., Hoque A.S.M.L (2015) Towards Development of National Health Data Warehouse for Knowledge Discovery, *Part of Advances in Intelligent Systems and Computing, vol 385. Springer, Cham*.

[277] SO Salinas, ACN Lemus (2019) Data warehouse and big data integration, *International Journal of Computer Science and Information Technology, Vol 9, No.2*.

[278] Hai, Rihan and Geisler, Sandra and Quix, Christoph (2016) Constance: An Intelligent Data Lake System, *Proceedings of the 2016 International Conference on Management of Data*.

The Use of Collaboration Distance in Scheduling Conference Talks

Jan Pisanski
University of Ljubljana, Faculty of Arts,
E-mail: Jan.Pisanski@ff.uni-lj.si, http://oddelki.ff.uni-lj.si/biblio/oddelek/osebje/pisanski.html

Tomaž Pisanski
University of Primorska, FAMNIT
E-mail: Tomaz.Pisanski@upr.si, https://en.wikipedia.org/wiki/Tomaz Pisanski
ORCID 0000-0002-1257-5376

Keywords: collaboration distance, scheduling, orthogonal partitions

Received: June 15, 2019

Several bibliographic databases offer a free tool that enables one to determine the collaboration distance or co-authorship distance between researchers. This paper addresses a real-life application of the collaboration distance. It concerns somewhat unusual clustering; namely clustering in which the average distances in each cluster need to be maximised. We briefly consider a pair of clusterings in which two cluster partitions are uniform and orthogonal in the sense that in each partition all clusters are of the same size and that no pair of elements belongs to the same cluster in both partitions. We consider different objective functions when calculating the score of the pair of orthogonal partitions. In this paper the Wiener index (a graph invariant, known in chemical graph theory) is used. The main application of our work is an algorithm for scheduling a series of parallel talks at a major conference.

Povzetek: Nekatere bibligrafske zbirke podatkov nudijo orodje, ki za poljubna raziskovalca poišče njuno razdaljo sodelovanja, oz. razdaljo soavtorstva. Članek obravnava konkretno uporabo razdalje sodelovanja. Pri tem gre za nekoliko nenavadno razvrščanje podatkov, pri katerem morajo biti razdalje med elementi skupine čim večje. Na kratko obravnavamo par uniformnih razvrščanj, pri katerem ima vsaka skupina prve komponente z vsako skupino druge komponente natanko en skupen element. Omenimo različne kriterijske funkcije za izračun vrednosti razvrščanj. V praksi uporabimo Wienerjev indeks, ki ga dobro poznamo v kemijski teoriji grafov. Glavna uporaba našega dela je algoritem za razporejanje serije vzporednih predavanj na večji konferenci.

1 Introduction

In this paper we address the use of collaboration distance in solving several practical problems. In particular we apply it to scheduling conference talks in parallel. A problem facing organizers of large conferences where several talks are scheduled in parallel is to avoid simultaneous talks of speakers that may interest the same person, or at least to minimize the number of attendees who have to choose between two interesting talks. Another, somehow complementary task is to schedule similar talks in the same session, preferably in the same lecture room and next to each other. So the main question is, what function one has to take to measure similarity between two speakers. In this paper we will use an objective approach to these ends and simply employ the collaboration distance, information that is readily available in some bibliographic databases.

2 Collaboration graph and collaboration distance

2.1 Collaboration graph

Let V be a list of researchers. This list may be obtained in any manner, but it makes sense to base it on (preferably authority controlled) lists of authors from bibliographic databases. We say that $u, v \in V$ are adjacent: $u \sim v$, if u and v collaborate. Usually, by collaboration we mean that they have written a joint publication in the past. In this sense we consider collaboration to be the same thing as co-autorship. Since $\sim$ is a binary irreflexive, symmetric relation it defines a simple graph $G = (V, \sim)$ that we call the *collaboration graph*[1]. Clearly, one has to specify the data set from which relation $\sim$ can be deduced. Hence G depends on the choice of such a data set.

[1]Here we present the basic model that suffices for our purposes. Note that some studies use reflexive relation signifying that each author collaborates with himself or herself. Also, the graph may be weighted where the weights on the edges represent the number of joint papers between the two authors.

2.2 Collaboration distance

Any connected graph G gives rise to a metric space where the distance $d(u, v)$ between two vertices $u, v \in V$ is defined as the length of the shortest path in G between u and v. If G is disconnected, each of its connected components is a metric space and we let $d(u, v) = \infty$ for vertices in different connected components. For a collaboration graph G the expression $d(u, v)$ is called a *collaboration distance* between authors u and v.

For basics in graph theory, the reader is referred to [6]; for metric spaces, see [7].

2.3 Data sets

It seems the first idea of collaboration graph and collaboration distance appeared as entertainment among mathematicians when measuring how close their research is from the prolific mathematician Pál Erdős. The corresponding collaboration distance is called the Erdős number, and was first formally introduced forty years ago [10]. Scientific investigation of Erdős collaboration graph began in 2000 [5]. Soon it became clear that the same data set can be used for computation of collaboration distance between any two individuals, not only the distance from one particular subject. One can easily define other collaboration graphs, e.g. among movie actors. There is an edge between two actors if and only if they have appeared in the same movie. Collaboration graphs became important in social sciences as prominent examples of social networks. Large social networks exhibit characteristic features of random networks. Modern theory of random networks was born in 1999 [1] when the model was proposed which explains very well the nature of social networks such as collaboration graphs.

Nowadays, two large bibliographic databases covering research in mathematics exist: *MathSciNet* that is run by the American Mathematical Society and *ZbMath*, run by the European Mathematical Society via Springer. Both cover most important publications in mathematics, statistics and theoretical computer science. Each of them contains a tool for calculating the collaboration distance between two authors. In our application the collaboration distances between speakers were taken from ZbMath.

Unfortunately, other important bibliographic databases such as Web of Science, SCOPUS or Google Scholar, do not provide free tools for computing collaboration distance. Slovenia has an excellent research information system SICRIS/COBISS that covers the work of over 15,000 Slovenian scientists. Although it has been analysed with respect to collaboration distance, only summary results in form of scientific papers are available, see e.g. [2, 3, 9, 11]. We strongly believe that a collaboration graph and the corresponding collaboration distance function based on SICRIS should be made available on-line.

3 Selecting optimal orthogonal partitions

Here we present an application of collaboration distance to a sample of individuals.

3.1 Scheduling talks in parallel

Let V be a set of speakers at a scientific conference. Assume each speaker delivers a single talk and that all talks are to be scheduled in parallel in m lecture rooms. Let $n = |V|$. To simplify our task we assume that there are t equal time-slots available and that $n = tm$.[2] Our task is to partition the set of speakers into t groups $U_1, U_2, \ldots, U_t$ such that each group U_i contains m speakers that will speak at the same time. At the same time we want to partition the speakers into m groups $L_!, L_2, \ldots L_m.$, assigning each group to a lecture room. In other words we are restricting our search to the pair of *uniform partitions*.

Group	L_1	...	$L_m.$
U_1	v_{11}	...	v_{1m}
U_2	v_{21}	...	v_{2m}
...	...	...	...
U_t	v_{t1}	...	v_{tm}

Table 1: Partitioning the set of speakers into t clusters U_i, representing time slots and an orthogonal partitioning into m clusters L_j , representing lecture rooms.

We would like to choose a partition in which the researchers in each part U work on different topics. A good measure may be collaboration distance.[3] If two researchers have a paper in common they should probably be in different parts. We would like collaboration distances in each group as big as possible. At the same time we would like to have the clusters in the other, orthogonal partition to be as homogeneous as possible. We decided to use a function that is well-known in chemical graph theory, namely, the *Wiener index*.

3.2 The Wiener index of an induced subgraph and clustering

Let G be a connected graph. The Wiener index $W(G)$ is defined as:

$$W(G) = (1/2) \sum_{u \in V} \sum_{v \in V} d(u, v)$$

[2] In more general case when the divisibility condition is violated one could introduce slack or dummy speakers and appropriately define the distances for them.

[3] Any of several other measures, such as citations, keywords, etc, could have been used.

We may restrict this index to a subgraph, induced by $U \subset V$.

$$W(G,U) = (1/2) \sum_{u \in U} \sum_{v \in U} d(u,v)$$

This notion can be found, for instance in [7].

Let $\mathcal{U}$ be a partition of V into t parts of size m, each. The partitioning may be called a *clustering* and each part may be called a *cluster*.

We generalise the notion of the transmission of a vertex in a graph; see [8]. Let v be a vertex, then the sum:

$$w(G,U,v) = \sum_{u \in U} d(v,u)$$

is called the *transmission* of v to U in G. Note that Dobrynin in [8] only considers the case when $U = V$. Given cluster U, the element $u \in U$ with minimal transmission is called a *clustroid* of U. Clustroids are used in several clustering algorithms. However, we will use them only *post festum*.

For a clustering $\mathcal{U}$ define

$$F(\mathcal{U}) = \sum_{U \in \mathcal{U}} W(G,U)$$

We are searching for an admissible partition $\mathcal{U}$ that will maximise $F(\mathcal{U})$. As we show below one may refine this search by adding another, orthogonal criterion.

3.3 Orthogonal clusters and orthogonal partitions

The same data and the same criterion function can be used in the opposite direction, namely to cluster speakers into sections. This means that the talks in the same section will be scheduled consecutively in the same lecture room.

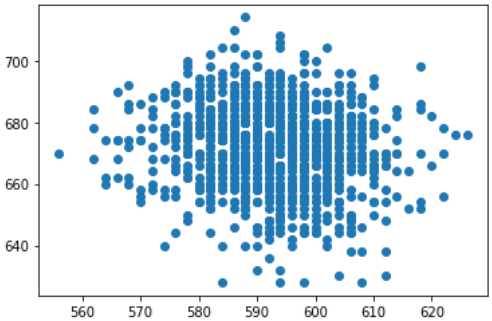


Figure 1: $F(\mathcal{U})$ vs. $F(\mathcal{L})$ for 10000 random permutations π. The optimal results and Pareto frontier can be found in the bottom right.

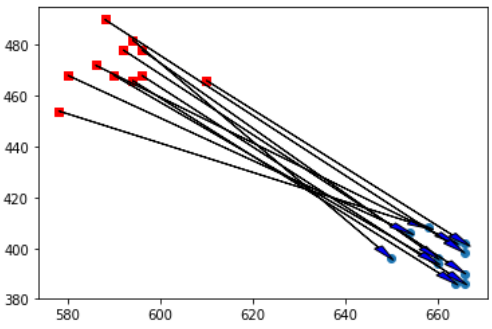


Figure 2: $F(\mathcal{U})$ vs. $F(\mathcal{L})$ for 10 random permutations π each followed by a local optimisation.The initial scores are in top left squares while the optimal scores and Pareto frontier can be found on the bottom right circles. Arrows link each square to the corresponding circle.

In case we want to perform both tasks simultaneously, we may choose to consider *orthogonal partitions*. Two uniform partitions of an mk-set are orthogonal if one has clusters of size k and the other one of size m and no pair of elements belongs to both partitions. In one partition we want to maximize distances while in its orthogonal mate we minimize distances.

Let $c = (\mathcal{U}, \mathcal{L})$ be a pair of orthogonal partitions of V. Let F be defined as above. Define the *score* of $(\mathcal{U}, \mathcal{L})$ to be $F(\mathcal{U}) - F(\mathcal{L})$. Note that each permutation π of V, i.e. $\pi \in Sym(V)$, can be considered as a pair $(\mathcal{U}, \mathcal{L})$. Hence $F(\pi) = F(\mathcal{U}) - F(\mathcal{L})$. We chose the solution to be $\mathrm{argmax}_{\pi \in \mathrm{Sym(V)}} \mathrm{F}(\pi)$.[4]

The task we wanted to solve was the scheduling of 30 invited speakers of the 8th European Congress of Mathematics that is taking place in Portorož, Slovenia in July 2020. The Congress takes place in 5 consecutive days and each day 6 speakers have to deliver their talks in parallel.

In the first attempt we generated 1000 admissible solutions randomly. The results are depicted in Figure 1. We also wrote a program for improving each admissible solution by local optimisation. This improved the quality of the final solution considerably. Figure 2 depicts 10 runs of our algorithm. The top left dots correspond to the randomly generated solutions while the bottom right ones depict the ones, obtained by a sequence of improvements leading to a local minimum. The arrows join each initial solution to the corresponding locally optimal one.

[4]Note that this can be considered also as a multi-criteria optimisation problem with score $(F(\mathcal{U}), -F(\mathcal{L}))$ with Pareto points being candidate solutions.

3.4 Alternative candidates for a score of an orthogonal pair of partitions.

The choice of $F(\pi)$ may not be most suitable for the task.

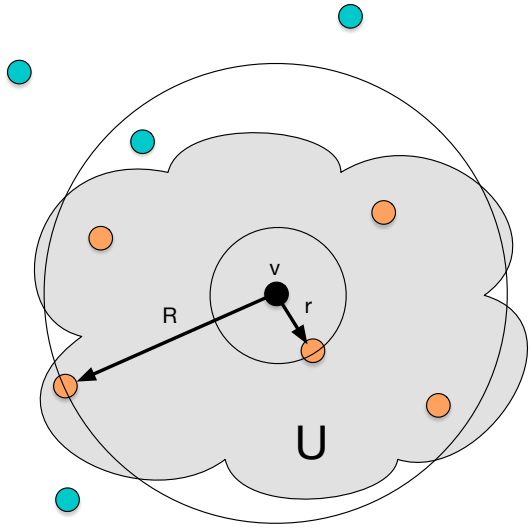


Figure 3: Radius R and the isolation radius r of point v with respect to U. Points in U are in the gray cloud.

We extend the definition of the *radius of a cluster* [12], to the radius of any individual v with respect to the cluster U:

$$R(G, U, v) = \max_{u \in U} d(v, u)$$

Since our clusters are in a sense *anticlusters* as they contain individuals being as far apart as possible, it make sense to define another radius that we call the *isolation radius*

$$r(G, U, v) = \min_{u \in U, d(u,v)>0} d(v, u)$$

measuring the distance to the nearest element in the cluster; see Figure 3.

Note that transmissions measure average distance, while the radius and isolation radius measure maximal and minimal distance, respectively. Also, the *centroid* is a vertex attaining the maximum radius in has been used extensively in data science. We may define *isolation centroid* as the vertex attaining minimal isolation radius.

Since we are already given a distance matrix, data preprocessing is not needed. If needed a method that has all clusters of equal size can be used.

It would be probably interesting to select the pair $(\mathcal{U}, \mathcal{L})$ by maximizing the sum of isolation radii in $\mathcal{U}$ and minimizing the sum of radii in $\mathcal{L}$. There are other well-known techniques, such as greedy method or integer programming that should be investigated for this problem.

4 Some further applications of collaboration distance

Collaboration distance can be used as a basis for natural structuring of a given list of researchers using standard clustering methods. We envision several applications of this approach including two that we mention here.

In the first approach one can focus on researchers belonging to a given organization, such as university, institute, faculty, department, project, etc. The internal structure of various universities and institutes could be compared to the collaboration network. Figure 4 is just an illustration of a simple application that gives a very natural stratification of a mathematical department in Slovenia in which three subgroups of researchers are clearly identified. Again, collaboration distances from ZbMath were used. We intend to pursue further studies in this direction.

The second one involves clustering of individuals of a given bibliographic database. Namely, having collaboration graph consisting of all researchers in a given database or country would be very useful. One could use it, in principle, to analyse similarity between various institutions, research groups and scientific disciplines. Various anomalies could be detected and used by policy makers to change the rules in order to avoid it in the future.

While the two mathematics databases (MathSciNet and ZbMath) provide the users with collaborative distance for a given pair of authors, most of the databases in other fields, as well as general databases, do not. This means that additional work must be done by users to find collaboration distances between authors. There are other factors to consider, when calculating Erdős numbers. Firstly, consistent data on the authors is needed, which implies at least consistent spelling of the names, but preferably authority control using consistent identifiers[5]. If this condition is not met, the results will not be appropriate. Next, the range of publications considered for calculation, can have a significant effect on the calculated collaboration distance. For a given pair of researchers their collaboration distance can be, and is, different for different databases. That only a certain subclass of publications is considered, is more or less an arbitrary decision, which is usually a reflection of the scope of a particular database.

One can envision other situations where different distances may be significant. For instance, when selecting referees for a paper one would like to select objective ones, i.e. the ones that are not co-authors of candidates. On the other hand we would like so select individuals who know well the subject, covered in the paper or project under review. This closeness may be measured, for instance by the overlap of keywords used by the two individuals.

Clearly, a *fractional approach* [4] in which the collaboration distance is not measured simply as a distance in the collaboration graph but the number of joint papers shortens the distance accordingly.

[5]One of the most well-know identifiers is ORCID.

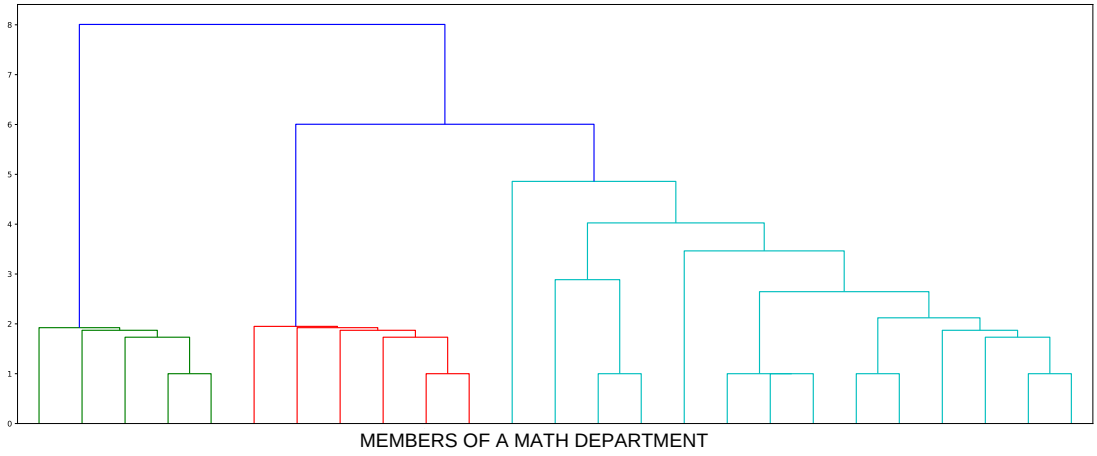


Figure 4: Ward method clustering based on the ZbMath collaboration distance for a department gives a reasonable partition of its members in three groups.

5 Conclusion

The main goal of this paper was to point out that the collaboration distance that is available at some high-quality bilbliographic databases such as MathSciNet and ZbMath is a useful tool that can be applied to a variety of specific problems such as scheduling talks at conferecnes or analysing internal structures of universities, institutes, etc. However, it would be very useful if one could specify the types of edges of the collaboration graphs. For instance, in MathSciNet co-authorship of editorial does not count. It would be useful if the user could choose criteria for inclusion/exclusion of data from the dataset. An important fact may be the time-frame of joint publications. For instance, by looking at recent co-authorships one could easily detect possible conflicts of interest. For other purposes it would be helpful to have information how many co-authors contributed to the edge of the collaboration graph and more generally the number of shortest paths connecting two authors.

Having such a simple tool incorporated into SICRIS would be an important upgrade of the system. One could also look at other measures of similarity, however, it would probably be difficult to get an agreement which ones to include. We would like to stress that we are not doing massive data mining. Our real-life calculations involved rather small data sets. For larger conferences with over 1000 active participants one should perhaps look for methods that would reduce the size of data that is needed to store the distance matrix. It would be interesting to explore how the attendees of a conference choose the talks they attend. In particular, it would be interesting to compare the proposed clustering approach to the manual organization of talks.

Acknowledgement

We would like to thank Vladimir Batagelj and Mark Pisanski for useful advice and fruitful discussion. The work of both referees is gratefully acknowledged. It improved the presentation of results considerably. Work of Jan Pisanski is supported in part by the ARRS grants P5-0361 and J5-8247, while work of Tomaž Pisanski is supported in part by the ARRS grants P1-0294,J1-7051, N1-0032, and J1-9187.

References

[1] A.L Barabási and R Albert. Emergence of scaling in random networks. *Science*, vol. 286 (1999), no. 5439, pp. 509–512. https://doi.org/10.1126/science.286.5439.509

[2] T. Bartol, K. Stopar, and G. Budimir. Visualization and knowledge discovery in metadata enriched aggregated data repositories harvesting from Scopus and Web of Science. *Information management in the big data era: for a better world : Selected IMCW2015 Papers.* Sun Yat-sen University North: Hacettepe University, 2015. pp 1–5.

[3] T. Bartol, et al. Mapping and classification of agriculture in Web of Science: other subject categories and research fields may benefit. *Scientometrics*, vol. 109 (2016), no. 2, pp. 979–996. https://doi.org/10.1007/s11192-016-2071-6

[4] V. Batagelj. On Fractional Approach to Analysis of Linked Networks, *arxiv* (2019) https://arxiv.org/abs/1903.00605.

[5] V. Batagelj and A. Mrvar. Some analyses of Erdős collaboration graph. *Social Networks,* vol. 22 (2000),

no. 2, pp. 173–186.
https://doi.org/10.1016/S0378-8733(00)00023-X

[6] J.A. Bondy and U.S.R. Murty. *Graph theory*, (2008) Graduate Texts in Mathematics, 244. Springer, New York.
https://doi.org/10.1007/978-1-84628-970-5

[7] M. M. Deza and E. Deza. *Encyclopedia of distances.* Fourth edition. (2016), Springer, Berlin.
https://doi.org/10.1007/978-3-662-52844-0

[8] A.A. Dobrynin. On 2-connected transmission irregular graphs *Diskretn. Anal. Issled. Oper.*, vol. 25 (2018), no. 4, pp. 5–14.

[9] A. Ferligoj et al. Scientific collaboration dynamics in a national scientific system. *Scientometrics*, vol. 104 (2015), no. 3, pp. 985–1012.
https://doi.org/10.1007/s11192-015-1585-7

[10] C. Goffman. And what is your Erdős number?, *Amer. Math. Monthly*, vol. 76 (1979), p. 791
https://doi.org/10.2307/2317868

[11] L. Kronegger, F. Mali, A. Ferligoj, and P. Doreian. Collaboration structures in Slovenian scientific communities. *Scientometrics*, vol. 90 (2012), no.2, pp. 631–647.
https://doi.org/10.1007/s11192-011-0493-8

[12] J. Leskovec, A. Rajaraman, and J. Ullman. *Mining of Massive Datasets* (2014), Cambridge University Press.
https://doi.org/10.1017/CBO9781139924801

[13] MathSciNet:

https://mathscinet.ams.org/mathscinet/index.html

[14] SICRIS:

https://www.sicris.si/public/jqm/cris.aspx?lang=eng

[15] zbMATH:

https://zbmath.org/

String Transformation Based Morphology Learning

László Kovács
Institute of Information Technology, University of Miskolc, Miskolc-Egyetemváros, H 3515, Hungary
E-mail: kovacs@iit.uni-miskolc.hu and https://www.iit.uni-miskolc.hu

Gábor Szabó
Institute of Information Technology, University of Miskolc, Miskolc-Egyetemváros, H 3515, Hungary
E-mail: szgabsz91@gmail.com and https://www.iit.uni-miskolc.hu

Keywords: machine learning, natural language processing, inflection rule induction, agglutination, dictionaries, finite state transducers, tree of aligned suffix rules, lattice algorithms, string transformations

Received: October 10, 2018

There are several morphological methods that can solve the morphological rule induction problem. For different languages this task represents different difficulty levels. In this paper we propose a novel method that can learn prefix, infix and suffix transformations as well. The test language is Hungarian (which is a morphologically complex Uralic language containing a high number of affix types and complex inflection rules), and we chose a previously generated word pair set of accusative case for evaluating the method, comparing its training time, memory requirements, average inflection time and correctness ratio with some of the most popular models like dictionaries, finite state transducers, the tree of aligned suffix rules and a lattice based method. We also provide multiple training and searching strategies, introducing parallelism and the concept of prefix trees to optimize the number of rules that need to be processed for each input word. This newly created novel method can be applied not only for morphology, but also for any problems in the field of bioinformatics and data mining that can benefit from string transformations learning.

Povzetek: Predstavljena je nova metoda za morfološko učenje na primeru madžarščine.

1 Introduction

In the area of natural language processing (NLP), word structure is an essential information for higher layer analysis such as syntax, part of speech tagging, named entity detection, sentiment and opinion analysis, and so on. The main difference between syntax and morphology is that while syntax works on the level of sentences, treating individual words as atoms, morphology works with intraword components.

According to morphology models, the words are built up using morphemes that are the smallest morphological units that encode semantic information. There are two types of morphemes: the lemma is the root, grammatically correct form of a word that's associated with the base meaning; while affixes are usually shorter character strings that slightly modify the meaning of the words. These affixes are language dependent, and can be prepended (***in****correct*), appended (*fly****ing***) or simply inserted into the words. Prepended affixes are called prefixes, appended affixes are called suffixes, while affixes inserted inside the words are called infixes. This latter category is rare in most languages, one example is the Latin verb *vi****n****cō* where the *n* denotes present tense. The addition of affixes is called inflection, while the inverse opereation is called lemmatization.

Languages can be categorized into six main groups based on their morphological features [1]. Analytic languages such as English have a fix set of possible affixes for each part-of-speech category. Isolating languages like Chinese and Vietnamese usually have words that are their own stems, without any affixes. Languages that have only a few affix types usually use auxiliary words and word position to encode grammatical information. In intraflective languages (Arabic, Hebrew), consonants express the meaning of words, while vowels add the grammatical meaning. Synthetic languages have three subcategories: polysynthetic languages like Native American languages contain complicated words that are equivalent to sentences of other languages; in fusional languages such as Russian, Polish, Slovak, Czech, the morphemes are not easily distinguishable and often multiple grammatical relations are fused into one affix; agglutinative languages like Hungarian, Finnish, Turkish have many affix types and each word can contain a large number of affixes.

For different languages there are different models that can be used to learn morphological rules, as morphology is a language dependent area. Creating such models is a complex task, especially for agglutinative languages. In the literature we can find approaches that are based on suffix trees and error-driven learning [2] to optimally store transformation rules and search among them.

Hajic [3] proposed a generalized grammar model,

suitable for both the synthetic and agglutinative languages. The author introduces a controlled rewriting system $CRS\,\langle A, V, K, t, R\rangle$, where A is the alphabet, V is the set of variables, K contains the grammatical meanings (morphological categories), t maps the variables to types and R is a set of atomic rewrite rules. The substitution operation defined in the rewrite rules replaces all variables with some string, all instances of the same variable is replaced by the same string. The main parameters of an elementary substitution rule include the input state id, the output state id, the variable id, the morphological category and the resulting string. The article provides a formal framework to describe the transformation process, but it does not detail the rule generation process, since the model assumes that the rule set is constructed by human experts.

In the two-level morphology model [4], the inflected words are represented on two levels. The outer or surface level contains the written form of the words, while the inner or lexical level contains the morphological structures. For example, the surface level word *"tries"* is related to the lexical level *"try+s"*. The lexical level represents the morphological categories and separator symbols for the surface form. The model uses a dictionary to store the valid lemmas and morpheme categories. The transformation between the lexical level and the surface level is implemented with a set of finite state transducers. A transducer is a special automaton that can model the string transformations.

FSTs (finite state transducers) are widely used to manage morphological analysis for both generation and recognition processes. One of the main issues related to this model is the computational complexity of the implementations. It was shown that it is inefficient to work with complex morphological constraints [5], where there are complex dependencies among the different morpheme units, like vowel harmony. The analysis shows that both recognition and generation are NP-hard problems. One of the most widely known approaches to construct an FST is the OSTIA method [6, 7]. It first generates a prefix tree transducer, then merges all the possible states, pushes some output elements toward the initial state and eliminates all the non-deterministic elements.

The OSTIA algorithm was later improved by Gildea and Jurafsky [8]. They extended the algorithm with a better similarity alignment component. Theron and Cloete [9] proposed a more general method based on edit-distance similarities of the base and inflected words. The algorithm learns the two-level transformation rules, calculating the string edit difference between each source-target pair and determining the edit sequences as a minimal acyclic finite state automaton. The constructed automaton can segment the target word into its constituent morphemes. The algorithm determines the minimal discerning context for each rule. This processing phase is done by comparing all the possible contiguous contexts to determine the shortest context.

Regarding current achievements, one important approach is presented in [10] and [11]. In the proposal of Goldsmith, a simplified morphology model is used containing substitution of suffixes. The words are decomposed into sets of short substrings, where the substrings have a role similar to the morphemes. The proposed method uses the concept of minimal description length to determine the appropriate word segmentations.

Another popular and simple method is the so-called tree of aligned suffix rules (TASR) [12] that is a great match for morphological rule induction: it can be built very quickly according to previous evaluations and can be searched very quickly as well, providing an outstanding correction ratio. Unlike dictionary based systems and FSTs, the TASR method can inflect even previously unseen words correctly. The only downside of this model is that it can only handle inflection rules that modify the end of the input word. In Hungarian we must be able to describe not only suffix rules, but also prefix and infix rules.

Besides trees, there are existing models that use lattice structures to store transformation rules. The goal of [13] is to optimize the lattice size by dropping rules that have a small impact on the overall results. The rule model uses similar concepts to the Levenshtein model like additions, removals and replacements. The paper shows that this lattice based model has a very promising memory constraint, fast inflection time and a correctness ratio of almost 100%.

In this paper we present a novel model called Atomic String Transformation Rule Assembler (ASTRA) whose base concept is similar to TASR, but can handle any types of affixes, including prefixes, infixes and suffixes as well. Our test language is Hungarian, a morphologically complex, highly agglutinative language that is frequently targeted by morphological model researchers due to its complexity. In Hungarian, there are a high number of affix types that can form long affix type chains, moreover each affix type can modify the base form significantly, using vowel gradation and changing consonant lengths. The inflection rules of the language are complex, and there are several exceptions, too. Besides morphological rule induction, our model is capable of dealing with any other string transformation based problems as well. Such problems can be found in the area of biological informatics (e.g. investigating DNA sequences) and data mining (e.g. preprocessing of data including spelling correction and data cleaning).

The structure of this paper is the following:

- Section 2 introduces the reference methods: dictionary based systems, finite state transducers, the tree of aligned suffix rules and the lattice based method.
- Section 3 describes the novel ASTRA method: its rule model, training phase and inflection phase. We also introduce three search algorithms to speed up inflection.
- The evaluation of the proposed method can be seen in section 4. The four metrics we measure and compare with the base methods are the training time, average inflection time, size and correctness ratio.

– In section 5 we present a general application of the ASTRA model.

2 Background

2.1 Dictionary based models

One of the most basic methods for learning inflection rules is using dictionaries. A dictionary can be considered as a $D \subseteq W \times W$ relation for morphological usage: for each input word it can return an output word.

Usually dictionaries not only contain the inflected forms of words, but also other semantic information like their meaning, part-of-speech tag, sample sentences and so on. There are many language dependent WordNet projects [14, 15] whose goal is to build such databases. Besides automatic data mining techniques, these databases are often validated and corrected by human experts.

Because of the large magnitude of data (the Hungarian WordNet contains more than 40,000 synsets, i.e. word sets with the same meaning), dictionaries can take much time to build. Their advantage is that irregular morphological forms are guaranteed to be retained, they aren't dropped by generalization techniques. Besides the training time, the downside of dictionaries is the lack of generalization: other automated methods usually can handle previously unseen words, too, but dictionaries can only inflect and lemmatize words they know.

2.2 Finite state transducers

Finite state automaton (FSA) is the base model for finite state transducers. An FSA is an $\mathcal{A} = \langle Q, \Sigma, q_\epsilon, E, F \rangle$ where Q is the finite set of states, Σ is the input alphabet, q_ϵ is the start state, $E : Q \times \Sigma \to Q$ is the state transition relation and F is the set of accepting states.

Finite state transducers (FST) [7] extend this model with additional components, as well as with outputting strings. There are multiple transducer models. A rational transducer is a $\mathcal{T} = \langle Q, \Sigma, \Gamma, q_\epsilon, E \rangle$ where Q, Σ and q_ϵ are the same as for an FSA; Γ is the output alphabet and $E \subset (Q \times \Sigma^* \times \Gamma^* \times Q)$ is the state transition relation. In practice, $\Sigma = \Gamma$. A sequential transducer is almost the same, except for two additional conditions: $E \subset (Q \times \Sigma \times \Gamma^* \times Q)$ and $\forall (q, a, u, q'), (q, a, v, q'') \in E \Rightarrow u = v$ and $q' = q''$. A subsequential transducer is a special sequential transducer that has a sixth component: $\sigma : Q \to \Gamma^*$ that is the state output function. Such transducer works in the following way: each input character causes a state transition and the label of this transition is appended to the output string. Finally, the ending state's output is also appended, resulting in the final output string. A transducer is onward if for every state, the state's output and the state transitions' outputs starting from this state have no common prefixes.

FSTs are used extensively while working with string transformations, because they have optimal sizes and can produce the output almost in constant time. However, as we'll see, with morphological applications, their generalization ability is not really usable.

2.3 Tree of aligned suffix rules

There are 3 main types of substrings that can change in a word during inflection: prefixes, suffixes and infixes. The substring pre $\in \Sigma^*, |\text{pre}| > 0$ is a prefix of the string $s_1 \in \Sigma^*$ if there exists another string $s_2 \in \Sigma^*$ such that $s_1 = \text{pre} + s_2$. Similarly, the substring suff $\in \Sigma^*, |\text{suff}| > 0$ is a suffix of the string s_1 if there exists another string s_2 such that $s_1 = s_2 + \text{suff}$. The substring inf $\in \Sigma^*, |\text{inf}| > 0$ is an infix of the string s_1 if there exist two other strings s_2, s_3 such that $s_1 = s_2 + \text{inf} + s_3$ where $|s_2| > 0$ and $|s_3| > 0$.

The TASR model can only work with morphological rules that modify the end of the words, meaning that it can only model suffix transformations. This restriction is acceptable for morphologically simpler languages, but complex agglutinative languages often contain prefix and infix transformation rules as well.

The goal of the TASR learning phase is to generate a set of suffix rules from a training word pair set. This set of rules is denoted by $\mathcal{R}_T = \{R_T\}$ in this paper. A suffix rule consists of two components: $R_T = (\sigma_T, \tau_T)$ where $\sigma_T, \tau_T \in \Sigma^*$. Here, σ_T contains the word-ending characters that are modified by the rule, and τ_T contains the replacement characters. As an example, for the English verb *try* whose past tense is *tried*, we can generate a suffix rule where $\sigma_T = \text{y}$ and $\tau_T = \text{ied}$.

The rule $R_{T_1} = \{\sigma_{T_1}, \tau_{T_1}\}$ is aligned with rule $R_{T_2} = \{\sigma_{T_2}, \tau_{T_2}\}$ or shortly $R_{T_1} \parallel R_{T_2}$ if $\forall s_1 \in \Sigma^* : \exists s_2 \in \Sigma^*$ such that $s_1 + \sigma_{T_1} = s_2 + \sigma_{T_2}$ and $s_1 + \tau_{T_1} = s_2 + \tau_{T_2}$. The aligned-with operator is symmetric, so $R_{T_1} \parallel R_{T_2} \iff R_{T_2} \parallel R_{T_1}$.

If we have a word pair, for example *(try, tried)* we can generate multiple aligned suffix rules. The minimal suffix rule is *(y, ied)*, and after extending this rule with one character at a time, we get *(ry, ried)* and *(try, tried)*.

We can define a frequency metric $\text{freq}(R_T \mid \mathbb{I})$ for each rule R_T based on the training word pair set $\mathbb{I} = \{(w_1, w_2) \mid w_1, w_2 \in W\}$, counting the number of word pairs for which R_T applies.

For every word pair in the training set, we must first generate all the aligned suffix rules according to the above definitions and insert these rules in a tree $(T, \subseteq)$. This tree will consist of nodes $n_{T_1}, n_{T_2}, \ldots, n_{T_m}$, each node n_{T_i} associated with a set of rules $n_{T_i} \mapsto \left\{R_{T_{ij}} = \left(\sigma_{T_{ij}}, \tau_{T_{ij}}\right)\right\}$. All the rules associated with the same node have the same context.

Let's have two nodes: $n_{T_\downarrow}$ and $n_{T_\uparrow}$. They are associated with the rules $R_{T_{\downarrow i}} = \left(\sigma_{T_{\downarrow i}}, \tau_{T_{\downarrow i}}\right)$ and $R_{T_{\uparrow j}} = \left(\sigma_{T_{\uparrow j}}, \tau_{T_{\uparrow j}}\right)$, respectively. The $n_{T_\downarrow}$ node is the child of $n_{T_\uparrow}$ or shortly $n_{T_\downarrow} \subset n_{T_\uparrow}$ if $\exists x \in \Sigma : \forall i, j : \sigma_{T_{\downarrow i}} = x + \sigma_{T_{\uparrow j}}$.

The root node and rules are denoted by $n_{T_\Uparrow} \mapsto \left\{R_{T_{\Uparrow k}} = \left(\sigma_{T_{\Uparrow k}}, \tau_{T_{\Uparrow k}}\right)\right\}$. For the root, the following condition applies: $\forall k : \left|\sigma_{T_{\Uparrow k}}\right| = \min_{ij} \left|\sigma_{T_{ij}}\right|$.

Child rule $R_{T_\downarrow} = \left\{\sigma_{T_\downarrow}, \tau_{T_\downarrow}\right\}$ is subsumed by parent rule $R_{T_\uparrow} = \left\{\sigma_{T_\uparrow}, \tau_{T_\uparrow}\right\}$ ($R_{T_\downarrow} < R_{T_\uparrow}$) if $\sigma_{T_\downarrow} = x + \sigma_{T_\uparrow}$ and $\tau_{T_\downarrow} = x + \tau_{T_\uparrow}$ where $x \in \Sigma$.

After these definitions, we can define which rule is the winning rule of node $n_{T_\downarrow}$ among the associated $R_{T_{\downarrow i}} = \left(\sigma_{T_{\downarrow i}}, \tau_{T_{\downarrow i}}\right)$ rules. Let $n_{T_\uparrow}$ be the parent node with rules $R_{T_{\uparrow j}} = \left(\sigma_{T_{\uparrow j}}, \tau_{T_{\uparrow j}}\right)$. The winner rule is $\hat{R}_{T_\downarrow} = R_{T_{\downarrow k}}$ such that $\text{freq}\left(R_{T_{\downarrow k}} \mid \mathbb{I}\right) = \max_i \left(\text{freq}\left(R_{T_{\downarrow i}} \mid \mathbb{I}\right)\right)$ and $\nexists j : R_{T_{\uparrow j}} > R_{T_{\downarrow k}}$.

After that we can build the tree from the generated rules. Typically the most general rules will be close to the root node, while the most specific rules will be stored in the leaves. Therefore, during inflection we can search the tree in a bottom-up fashion, returning the winner rule of the first node we find whose context matches the input word. Since we start at the leaves, the first matching rule will be the most specific one, having the longest context. This means that the resulting inflected form will mirror the main characteristics of the training data.

2.4 Lattice based method

The rule model of the examined lattice based inflection method [13] is a six-tuple $R = (\alpha, \sigma, \omega, \overrightarrow{\eta}, \overleftarrow{\eta}, \langle\delta_i\rangle)$, where

- $\alpha \in \Sigma^*$ is the prefix of the rule containing the characters before the changing part,
- $\sigma \in \Sigma^*$ is the core of the rule that is the changing part,
- $\omega \in \Sigma^*$ is the postfix of the rule containing the characters after the changing part,
- $\overrightarrow{\eta} \in \mathbb{N}$ is the front index of the rule's context occurrence in the source word,
- $\overleftarrow{\eta} \in \mathbb{N}$ is the back index of the rule's context occurrence in the source word and
- $\langle\delta_i\rangle$ is a list of simple transformation steps on the core, $\delta_i \subseteq \Sigma \times \Sigma$.

These rules are generated automatically from training word pairs, then inserted into a lattice structure, where the parent-child relationship is based on rule context containment. In the original paper we formalized multiple lattice builder algorithms that tried to reduce the size of the resulting lattice. The best builder only inserts those rules and intersections into the lattice that are really responsible for the high correctness ratio, every other redundant rule is eliminated.

As we'll see, the size characteristics of this model is very promising, but because of the high degree of generalization, the lattice can inflect some words incorrectly. This is due to the overgeneralization effect of the lattice model itself.

3 Atomic string transformation rule assembler

The goal of the Atomic String Transformation Rule Assembler (ASTRA) model is to collect atomic, elementary patterns from a training word pair set during the training phase, and use the best matching atomic rules for each input word during the production phase. For these inputs, every matching, non-overlapping atomic rule is applied to produce the correct inflected form. As discussed previously, using these concepts, the proposed method can model prefix, infix and suffix inflection rules as well, thus can be used for morphologically complex agglutinative languages.

First of all, we define an extended alphabet so that it is easier to determine where a word starts and ends. Let's introduce two special characters, \$ that will mark the start of the word and # that will mark the end of the word. If a rule's context contains any of these two special characters, it will be easier to determine if the beginning or the end of the word needs to be transformed.

Of course these characters are not part of the original Σ alphabet. The extended alphabet will be denoted by $\bar{\Sigma} = \Sigma \cup \{\$, \#\}$. We also define a new operator on strings that prepends \$ and appends # to the string s: $\mu(w) = \bar{w} = \$ + w + \#$. The inverse operation drops the special characters from the input word: $\mu^{-1}(\bar{w}) = w$. The set of extended words is denoted by $\bar{W}$.

The input of the training process for the new method is the same set of word pairs containing the base form and inflected form of the word, but the first step of the algorithm is to extend these word pairs with our new special characters. After the extension, we get a new training set $\bar{\mathbb{I}} = \{(\bar{w}_1, \bar{w}_2)\}$.

We split each word pair to matching segments

$$\bar{w}_1 = \psi_1^1 \psi_1^2 \ldots \psi_1^k$$
$$\bar{w}_2 = \psi_2^1 \psi_2^2 \ldots \psi_2^k$$

A segment $\psi_1^i \to \psi_2^i$ is called variant if $\psi_1^i \neq \psi_2^i$, otherwise it is called invariant. In a segment decomposition, variant and invariant segments are alternating.

As one word pair might have multiple segment decompositions, we need to select the best one among them. To quantify the goodness of the decompositions, we use a segment fitness formula that returns how well-aligned the $\psi_1^i \to \psi_2^i$ segment is:

$$\lambda_1 \cdot \frac{1}{\text{index}_{\text{max}} - \text{index}_{\text{min}}} + \lambda_2 \cdot \left|\psi_2^i\right|$$

where $\text{index}_{\text{max}}$ and index_{min} are the maximal and minimal indices of the ith segment, i.e. the maximum and minimum of the indices $\sum_{j=1}^{i-1} \left|\psi_1^j\right|$ and $\sum_{j=1}^{i-1} \left|\psi_2^j\right|$, respectively. This formula encodes that invariant segments are better if their components are longer and the two components appear near to each other.

Example 3.1. *Let us choose a training word pair (dob, ledobott)*[1] *as an example to demonstrate the segment decomposition algorithm. First, the words are extended with the special characters: (\$dob#, \$ledobott#). One valid segment decomposition is the following: (ψ_1^1 = \$, ψ_2^1 = \$le), (ψ_1^2 = dob, ψ_2^2 = dob), (ψ_1^3 = #, ψ_2^3 = ott#). The middle segment is invariant, while the first and last ones are variant segments.*

For each variant segment, we can define so-called atomic rules in the form of $R_A = (\alpha_A, \sigma_A, \tau_A, \omega_A)$ where α_A is the prefix and ω_A is the suffix. The rule context that must be searched in the input words later is $\gamma_A(R_A) = \alpha_A + \sigma_A + \omega_A$. We can see that with this rule model, not only suffix rules can be modelled, because of the new α_A and ω_A components.

Let's take a variant segment $\psi_1^i \rightarrow \psi_2^i$. First, we need to define the core atomic rule $R_{A_{ic}} = (\alpha_{A_{ic}}, \sigma_{A_{ic}}, \tau_{A_{ic}}, \omega_{A_{ic}})$ for this segment that has no prefix or postfix, i.e. $|\alpha_{A_{ic}}| = 0$, $\sigma_{A_{ic}} = \psi_1^i$, $\tau_{A_{ic}} = \psi_2^i$ and $|\omega_{A_{ic}}| = 0$.

Then, we can extend this core atomic rule with one character at a time on the left and right sides, symmetrically. Let's assume that $\sum_{j=1}^{i-1} \left|\psi_1^j\right| = n$, $\sum_{j=i+1}^{k} \left|\psi_1^j\right| = m$ and $\left|\psi_1^i\right| = l$. In this case, the extended rule candidates are $R_{A_{ij}} = \left(\alpha_{A_{ij}}, \sigma_{A_{ij}}, \tau_{A_{ij}}, \omega_{A_{ij}}\right)$ with the following components ($\forall 1 \leq j \leq \min\{n, m\}$):

$$\begin{aligned}
\alpha_{A_{ij}} &= \bar{w}_1[n+1-j,\ n] \\
\sigma_{A_{ij}} &= \psi_1^i \\
\tau_{A_{ij}} &= \psi_2^i \\
\omega_{A_{ij}} &= \bar{w}_1[n+l+1,\ n+l+j]
\end{aligned}$$

Here, $w[i,j]$ denotes the substring of w from the ith to the jth character.

To make the generated atomic rules unambiguous, we have to make sure that the context of the rules only appear once in the base form of the word ($\bar{w}_1$). Every atomic rule candidate whose context appears more than once in the base form of the word is dropped from the final set.

Example 3.2. *Using the winning segmentation of example 3.1, the following atomic rules can be generated from the word pair (dob, ledobott):* $(-, \$, \$le, -)$, $(-, \$, \$le, d)$, $(-, \$, \$le, do)$, $(-, \$, \$le, dob)$, $(-, \$, \$le, dob\#)$, $(-, \#, ott\#, -)$, $(b, \#, ott\#, -)$, $(ob, \#, ott\#, -)$, $(dob, \#, ott\#, -)$, $(\$dob, \#, ott\#, -)$.

Transforming a word $\bar{w} \in \bar{W}$ using the atomic rule $R_A = (\alpha_A, \sigma_A, \tau_A, \omega_A)$ can be defined as

$$\chi_A(R_A, \bar{w}) = \begin{cases} \bar{w} \text{ if } \gamma_A(R_A) \not\subseteq \bar{w}, \text{ or} \\ \bar{w} \setminus \gamma_A(R_A)[\sigma_A \rightarrow \tau_A] \end{cases}$$

where $\bar{w} \setminus \gamma_A(R_A)[\sigma_A \rightarrow \tau_A]$ means that we need to search $\gamma_A(R_A)$ in $\bar{w}$, and replace σ_A with τ_A.

[1]Hungarian for *(throw, threw down)*. Note that we add two affixes, one for the past tense and one preverb for *down*.

The base form of the method doesn't require to build a tree, we can simply group the atomic rules based on their contexts. A rule group is defined as a set of atomic rules $\Gamma_A = \{R_{A_i} = (\alpha_{A_i}, \sigma_{A_i}, \tau_{A_i}, \omega_{A_i})\}$ where $\forall R_{A_i}, R_{A_j} \in \Gamma_A : \gamma_A(R_{A_i}) = \gamma_A\left(R_{A_j}\right)$. The context of the rule group is $\gamma_A(\Gamma_A) = \gamma_A(R_A)\ \forall R_A \in \Gamma_A$.

Example 3.3. *For the atomic rules of example 3.2, we can produce nine different rule groups, each containing a single atomic rule except for the rule group with context \$dob# that contains both* $(-, \$, \$le, dob\#)$ *and* $(\$dob, \#, ott\#, -)$.

The goal of the training phase is to produce a set of rule groups $\mathcal{R}_A = \{\Gamma_A\}$ based on the training word pair set $\bar{\mathbb{I}}$. The generated atomic rule set can be used to inflect the given input words based on the training word pair set. For each input, our goal is to choose some atomic rules that match the input word. Rules with longer matching substrings in the input word are better than rules with shorter matching substrings. The fitness function is

$$f(R_A \mid \bar{w}) = \frac{|\gamma(R_A)|}{|\bar{w}|} \cdot \theta^k(\gamma(R_A), \bar{w})$$

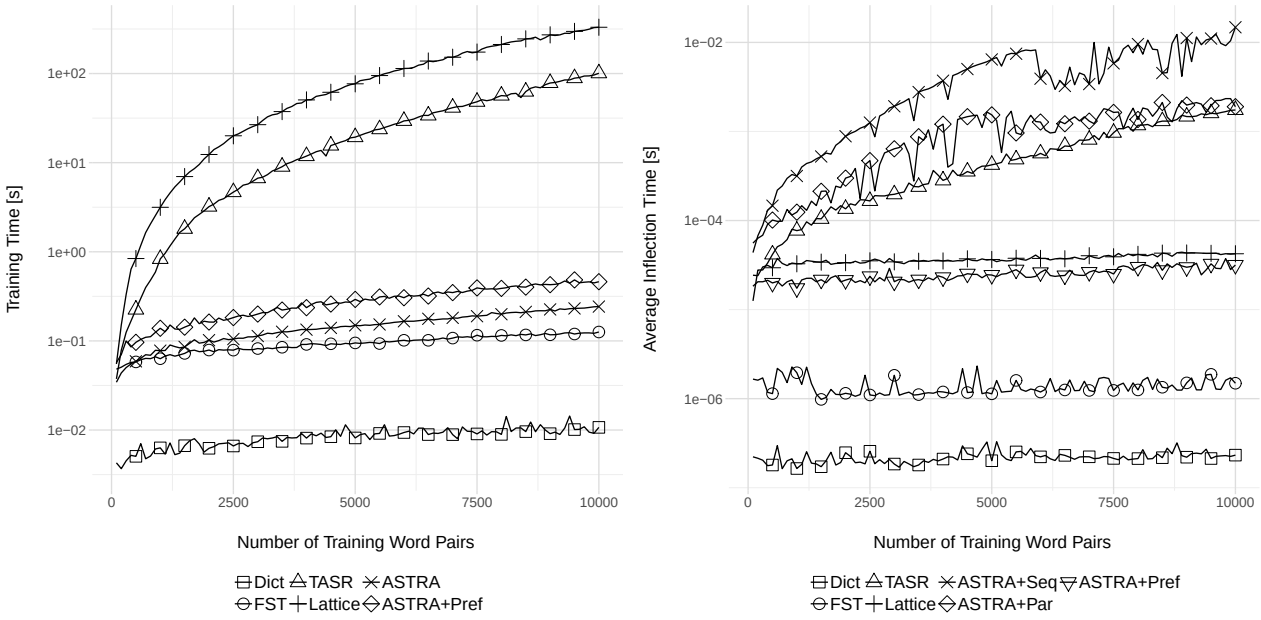
where k is a parameter and the θ function returns how similar the rule context is to the input word. To simplify things, we used $k = 1$ and a discrete θ function that returns 1 if $\gamma(R_A) \subseteq \bar{w}$, and 0 otherwise.

Using this fitness function, we can choose the first n atomic rules that are best suited for the given input word where n is a parameter. We implemented three separate candidate selector algorithms. The first one is a sequential algorithm that processes each rule group one by one. If a rule group's context matches the input word, its atomic rules are added to the resulting set of candidate rules. The second one is a parallel algorithm that does the same thing in a divide and conquer manner, processing the rule groups in parallel. The number of threads depends on the number of our CPU cores. The third one uses a prefix tree that is built from the rule groups during the training phase. With the prefix tree, we can speed up the candidate search process by searching substrings of the input words. If a substring is found in the prefix tree, the appropriate rule group's atomic rules are added to the resulting set.

Since there might be multiple overlapping rule candidates that would transform the same substring of the word leading to ambiguity, among these rules only the first one is used, the others are dropped. After we chose the best non-overlapping rules, we can apply them one by one on the input word, producing its inflected form.

4 Evaluation of the proposed method

For evaluation purposes, we used a training word pair set generated by [16]. We chose the Hungarian accusative case



(a) Training time of the methods

(b) Average inflection time of the methods

Figure 1: Training time and average inflection time of the methods

as our target affix type and used up to 10,000 training word pairs.

We compared a custom dictionary implementation, Lucene's FST method, the TASR model, the previously mentioned lattice based method and the proposed ASTRA method, measuring their training times, their average inflection times, the sizes of their rule base and their correctness ratios, i.e. how much percent of evaluation words are inflected correctly after the training phase. If W^+ is the set of evaluation words for which the model yields a correct inflected form, and W^- is the set of failed evaluation words, then the correctness ratio is $W^+ / (W^+ + W^-)$. Where applicable, we also measured the differences using the sequential, parallel or prefix tree search algorithm in case of ASTRA.

In Figure 1a we can see the training time of the methods, using logarithmic scale for the y axis. As we can see, there are three different clusters based on the training time. The fastest solution is to store the already available set of word pairs in a dictionary, because we only have to store these records, no extra processing occurs. Building an FST is the next in line, but it has very similar characteristics to the ASTRA method. If we include the prefix tree building as well, the ASTRA's training time increases a bit. The third cluster consists of the TASR and the lattice based methods. It can be seen that building a tree of aligned suffix rules takes more time as the previous methods, and the complexity of the lattice adds even more time to the TASR's results.

Figure 1b shows the average inflection time of the methods. As we can expect, if we use an appropriate hash function in the dictionary implementation, retrieving the matching record for each input word becomes almost constant in time. The second best method as for average inflection time is the FST: it also has a very plain curve, but it's a bit higher than the dictionary's. ASTRA with a prefix tree comes next, but it's very close to the line of the lattice based method. The remaining methods have much steeper curves: TASR comes next, but the parallel search function with ASTRA is very close to it; while the worst inflection time is achieved by the sequential search function. Note that although the inflection time of the prefix tree search variant is the best for ASTRA, it means a bit overhead during the training time. However, even with this overhead, we can say that it's worth using it.

In Figure 2 we can see the overall size of the rule bases, i.e. the number of word pairs in the dictionary, states in the FST, nodes in the TASR and the lattice, and atomic rules in case of ASTRA.

It is not surprising that there are more generated atomic rules in ASTRA than nodes in the tree of aligned suffix rules, since the atomic rule definition allows to have multiple variant segments in a word pair and from these variant segments, multiple core and extended atomic rules can be produced. On the other hand, TASR will only generate one minimal suffix rule per word pair and all of its aligned extensions. The advantage of the ASTRA model is that even with this higher number of rules and the prefix tree, we can train it faster than a TASR. Moreover it can cover more cases, including prefix, infix and suffix rules. The built FST has better size characteristics, because its builder algorithm merges every state that can be merged without losing information from the original training word pair set. It can be seen from the line of the dictionary that the number of states in an FST and the number of rules in the AS-

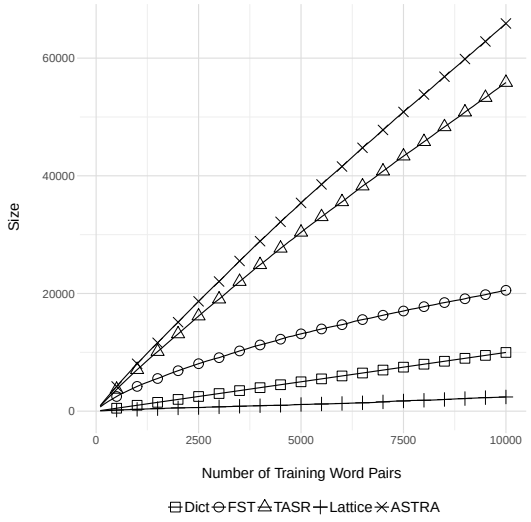


Figure 2: Size of the rule bases

TRA and TASR are higher than the number of input word pairs. However, the minimal lattice builder algorithm produces an even better lattice size, as the number of nodes in the resulting lattice is lower than the size of all the other structures.

Finally, Figures 3a and 3b show the correctness ratio of the models. The results of the left side were achieved by using disjoint training and evaluation word pair sets. We can see that the correctness ratio platoes a bit below 95% for TASR and ASTRA, the latter one performing a bit better. It can be also seen that the lattice based method is worse, probably because of its higher degree of generalization. When we examined the results of the lattice compared to TASR and ASTRA, we saw that in multiple cases the lattice found a node whose rule resulted in an invalid inflected form. The correctness ratio of the dictionary and the FST is 0%, because they could not generalize at all. For the dictionary, it is understandable, because a dictionary is a static map of word pairs. On the other hand, although an FST can generalize, these types of morphological applications don't benefit from this generalization, as the generalized transformations do not result in real inflection rules.

On the right side of the figure, we can see what happens if we use the first 100, 200, ... 10,000 word pairs to train the methods, and then use the same 10,000 word pairs for evaluation. All the methods have an almost 100% correctness ratio at the end of the diagram. The only reason that we cannot reach 100% is that in the training word pair set there are records with the same lemma and different inflected forms such as *örömöt* and *örömet* that are two valid inflected forms of the Hungarian word *öröm* (*joy* in English). The difference resides in the characteristics of the curves. The dictionary and the FST cannot really generalize inflection rules, so their lines are linear. The other methods can reach higher percentages more quickly, but as we can see, the ASTRA method is even better than the TASR in that it can produce a better correctness ratio with a smaller number of training word pairs. The lattice based method is worse than TASR and ASTRA in this case as well.

5 General application of the ASTRA model

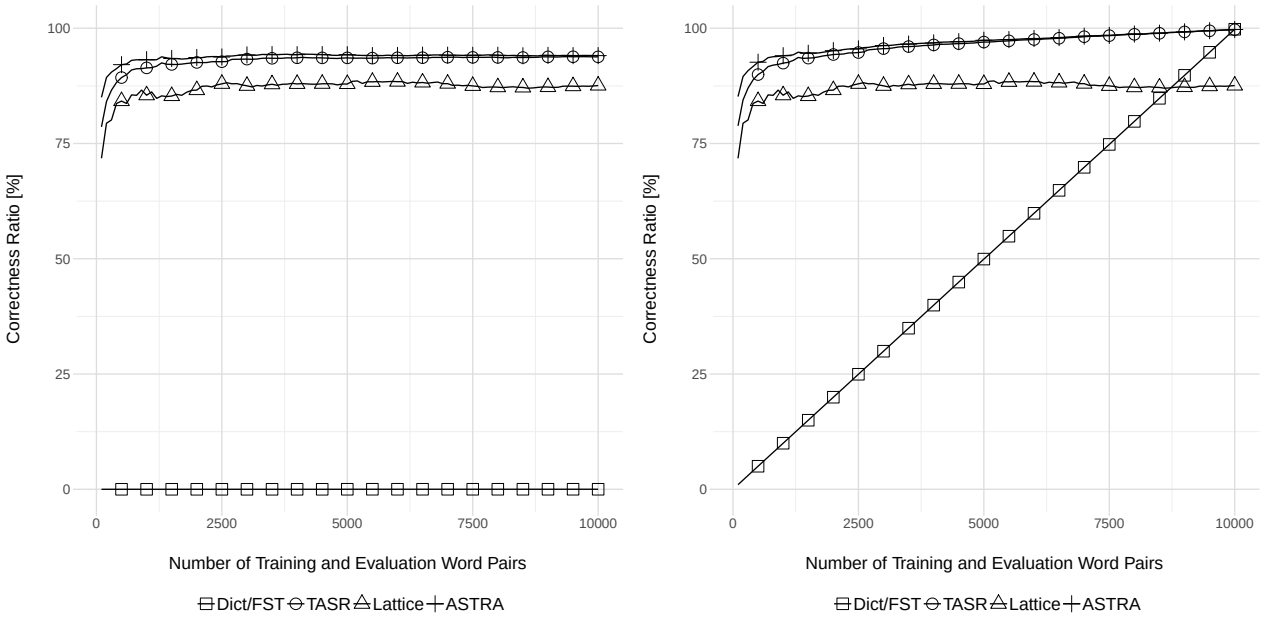
One of the scientific areas of applying string algorithms including string transformation based methods is the area of bioinformatics and computational biology [17]. DNA sequences are modelled using strings of four characters matching the four types of bases: adenine (A), thymine (T), guanine (G) and cytosine (C). One of the goals of bioinformaics is to compare genes in DNAs to find regions that are important, find out which region is responsible for what functions and features and determine how genetic information is encoded. The process of DNA analysis is a very computational intensive task, that's why modelling, statistical algorithms and mathematical techniques are important aspects of success.

Besides applying string transformations, computational biology uses many string matching and comparison techniques as well [18]. Finding the longest matching substrings of two strings (DNA sequences) helps in finding the best DNA alignments and thus comparing different DNA sequences, finding matching parts and differences. One of the techniques used for this comparison is the application of the edit distance computation originally published by Levenshtein [19] for morphological analysis.

Another application area where string transformation based methods are applied is data mining. Data mining engines usually consist of multiple phases to extract information out of unannotated training data such as long free texts. The first phase is often called data cleaning, where the raw input data is preprocessed so that invalid records are either removed or fixed before moving on with the data mining algorithms. One way to fix the typos and other errors in free texts is spelling correction. Spelling correction can be interpreted as learning those string transformations that can transform an unknown word containing typos to the closest known word. There are multiple techniques to solve this problem, usually iterative algorithms perform better as there can be multiple problems with a word that are easier to fix in multiple steps [20]. The goal is to find a word $w \in W$ for any unknown string s so that their distance $d(w, s) < \delta$ is lower than an acceptable threshold.

A third, more intuitive non-morphological application of the ASTRA model is character sorting. Let's have a random string $s \in \Sigma^*$ with a given length of $|s| = n$. The goal is to rearrange the characters in s so that for each index i, $1 \leq i < |s|$, $s_i \leq s_{i+1}$ for a given partial ordering, for example lexicographic ordering.

For our evaluation, we used input lengths 100, 200, ..., 3000. For each input length, we generated a random string and applied a pre-trained ASTRA on the string incrementally until the output was equal to the input. Then we



(a) Disjoint training and evaluation data sets

(b) Common training and evaluation data set

Figure 3: Correctness ratio with disjoint and common training and evaluation data sets

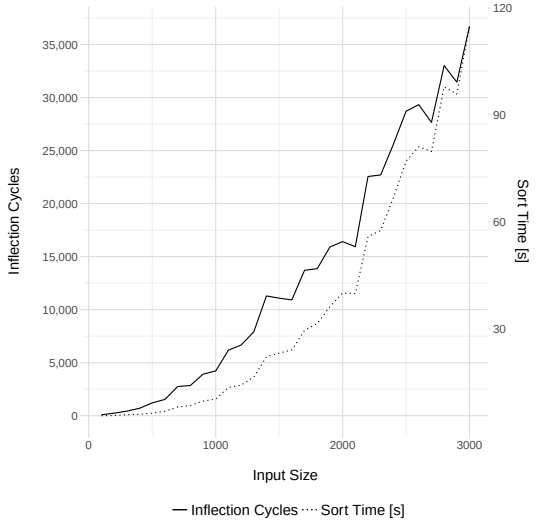


Figure 4: Inflection Cycles

checked if the final output contained the expected ordering, and found that all of the results were correct.

For the training process of the ASTRA, we generated a training word pair set. Each word pair contained a necessary transformation as the core, such as *(ba, ab)*, *(ca, ac)*, ..., *(zy, yz)*. To make the rules more noisy, we also generated a random string of 10 characters and prepended and appended it to both words in the word pair. For each word pair, this random prefix-suffix part was different. The results were all correct. The number of required iterations and the sorting time is displayed in Figure 4.

Unlike ASTRA, the other examined methods could not sort the characters correctly. The dictionary and FST methods, as we saw previously, cannot be used for inputs that are not present in the training word pairs set. TASR can only transform inputs that should be modified at the end. The lattice based method's disadvantage in this case is that it is not position agnostic, therefore it cannot determine the atomic transformations necessary for sorting the characters.

6 Conclusion

In this paper we presented the novel ASTRA model. The motivation was that although the TASR method can handle suffix morphological rules extremely well, it cannot describe rules modifying the beginning or the middle of words. In the target language of our research, Hungarian, there are a few affix types that have prefix inflection rules. The proposed rule model contains multiple components to not only store the changing part of the word, but also its preceding and following characters. We also defined a novel training algorithm that can generate such rules and store them in rule groups. A fitness function was defined that helps us choose the best rules from the rule database for each input word and make sure we can produce the inflected form easily. Finally, we implemented three search algorithms: one sequential, one parallel and one prefix tree based search function. We evaluated the proposed method, comparing its training time, average inflection time, size and correctness ratio with the same metrics of some base models, including a dictionary based system, Lucene's FST implementation, the TASR method and a lattice based model. The training time of ASTRA is ex-

ceptional, only the dictionary's and FST's training times are better, even if we also build a prefix tree from the generated rules. The same can be said about the average inflection times. The size of ASTRA is the worst compared to the other methods, but this is not really a problem, because the inflection time does not get worse, and we can handle more general inflection rules. The correctness ratio is also exceptional, moreover it reaches higher percentages even with less knowledge, i.e. fewer training word pairs than for example the TASR method. Besides these metrics, the advantage of the proposed novel ASTRA method is that it can be used not only for morphological rule induction, but also for any types of problems that can be modelled with string transformations. To demonstrate this, we adapted ASTRA to a character sorting problem with a correction ratio of 100%.

Aknowledgement

The described article/presentation/study was carried out as part of the EFOP-3.6.1-16-00011 "Younger and Renewing University - Innovative Knowledge City - institutional development of the University of Miskolc aiming at intelligent specialisation" project implemented in the framework of the Szechenyi 2020 program. The realization of this project is supported by the European Union, co-financed by the European Social Fund.

References

[1] A. Gelbukh, M. Alexandrov and S.-Y. Han (2004) Detecting inflection patterns in natural language by minimization of morphological model, *Iberoamerican Congress on Pattern Recognition*, Springer, pp. 432—438. https://doi.org/10.1007/978-3-540-30463-0_54

[2] G. Satta and J. C. Henderson (1997) String transformation learning, *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, pp. 444-–451.

[3] J. Hajic (1988) Formal morphology, *In Proceedings of International Conference on Computational Linguistics*, pp. 223–229. https://doi.org/10.3115/991635.991680

[4] K. Koskenniemi (1983) *Two-level morphology: A General Computational Model for Word-Form Recognition and Production*, Department of General Linguistics, University of Helsinki, Finland.

[5] L. Bauer (2003) *Introducing linguistic morphology*, Edinburgh University Press.

[6] J. Oncina, P. García and E. Vidal (1993) Learning subsequential transducers for pattern recognition interpretation tasks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 15, Number 5, pp. 448–458. https://doi.org/10.1109/34.211465

[7] C. De la Higuera (2010) *Grammatical inference: learning automata and grammars*, Cambridge University Press. https://doi.org/10.1017/CBO9781139194655

[8] D. Gildea and D. Jurafsky (1995) Automatic Induction of Finite State Transducers for Simple Phonological Rules, *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, Cambridge, Massachusetts, pp. 9–15.

[9] P. Theron and I. Cloete (1997) Automatic acquisition of two-level morphological rules, *Proceedings of the fifth conference on Applied natural language processing*, pp. 103–110.

[10] J. Goldsmith (2006) An algorithm for the unsupervised learning of morphology, *Natural Language Engineering*, Volume 12, Number 4, pp. 353–371. https://doi.org/10.1017/S1351324905004055

[11] J. Lee and J. Goldsmith (2016) Linguistica 5: Unsupervised Learning of Linguistic Structure, *HLT-NAACL Demos*, pp. 22–26.

[12] K. Shalonova and P. Flach (2007) Morphology learning using tree of aligned suffix rules, *In ICML Workshop: Challenges and Applications of Grammar Induction*.

[13] G. Szabó and L. Kovács (2018) Lattice based morphological rule induction, *Acta Universitas Apulensis*, Number 53, pp. 93–110. https://doi.org/10.17114/j.aua.2018.53.07

[14] C. Fellbaum (1998) *WordNet: An electronic lexical database*, MIT Press, Cambridge.

[15] M. Miháltz, Cs. Hatvani, J. Kuti, Gy. Szarvas, J. Csirik, G. Prószéky and T. Váradi (2007) Methods and results of the Hungarian WordNet project, *In Proceedings of GWC 2008: 4th Global WordNet Conference*, University of Szeged, pp. 311–320.

[16] G. Szabó and L. Kovács (2015) Efficiency analysis of inflection rule induction, *In Proceedings of the 2015 16th International Carpathian Control Conference (ICCC)*, IEEE, pp. 521-–525.

[17] N. C. Jones (2004) *An introduction to bioinformatics algorithms*, MIT press.

[18] E. Mourad and Z. Y. Albert (2011) *Algorithms in computational molecular biology: techniques, approaches and applications*, Volume 21, John Wiley & Sons.

[19] V. I. Levenshtein (1966) Binary codes capable of correcting deletions, insertions, and reversals, *Soviet physics doklady*, Volume 10, Number 8, pp. 707–710.

[20] S. Cucerzan and E. Brill (2004) Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users, *EMNLP*, Volume 4, pp. 293–300.

Feature Augmentation Based Hybrid Collaborative Filtering Using Tree Boosted Ensemble

Udayabalan Balasingam and Gopalan Palaniswamy
Department of Computer Applications, National Institute of Technology, Tiruchirappalli, India
E-mail: udayabalan@yahoo.com, npgopalan@nitt.edu

Keywords: feature augmentation, collaborative filtering, ensemble modelling, boosting, decision tree, supervised learning

Received: January 8, 2018

Requirements for recommendation systems are currently on the raise due to the huge information content available online and the inability of users to manually filter required data. This paper proposes a Feature augmentation based hybrid collaborative filtering using Tree Boosted Ensemble (TBE), for prediction. The proposed TBE recommender is formulated in two phases. The first phase creates category based training matrix using similar user profiles, while the second phase employs the boosted tree based model to predict ratings for the items. A threshold based filtering is finally applied to obtain precise recommendations for the user. Experiments were conducted with MovieLens dataset and performances were measured in terms of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The proposed model was observed to exhibit MAE levels of 0.64 and RMSE levels of 0.77 with a variation level of ±0.1. Comparisons with state-of-the-art models indicate that the proposed TBE model exhibits reductions in MAE at 6% to 14% and RMSE at ~0.2.

Povzetek: Avtorji so razvili novo metodo za svetovalne sisteme, temelječo na bogatitvi atributov za filtriranje s pomočjo spodbujevalnega ansambla dreves (TBE).

1 Introduction

Information explosion has led to a huge amount of data being generated online. However, human intellect and perception levels are stable, leading to difficulty in processing all the information available to them [1]. This has led to the formulation of prescription based models that provides automatic recommendations to the users. Such automated recommendations help users to a large extent by categorizing the information and providing the most significant information to the users such that they do not miss them. Systems enabling such automated categorizations and filtering are called recommender systems.

Recommender or a recommendation system is a specialized information filtering model that performs predictions based on the preference a user provides to an item. The preference levels are measured using ratings provided by the user to the item or similar items [2]. User ratings are analyzed and items similar to best rated items are recommended for the users. Recommender systems are not sidelined to predicting products alone. Due to the high online usage levels, such systems have become very popular and are currently used to predict books [3], music, news, research articles and even search queries, jokes and restaurants. Some of the current and most popular recommendation systems were music predictions by Last.fm and Pandora radio. Interests in recommendation systems were sparked by the Netflix challenge that offered a prize of $1 Million for improving their model by 10% [4].

Recommendation systems can be designed in three major aspects [5] namely; collaborative filtering, content based filtering and hybrid recommenders. Collaborative filtering models [6] are based on analyzing the user's behaviors and preferences to provide predictions. Major advantages of using such models are that they are based on available and machine analyzable content. This makes their recommendations more accurate and relatable. However, they suffer from issues like data sparsity, data unavailability (cold start) [7] and data volume [8]. Content based recommenders [9] are based on items that model user's profiles. Predictions are based on the created profiles. Hybrid recommenders [10] are a combination of collaborative and content based recommenders.

This paper proposes a feature augmentation based collaborative filtering mechanism to predict items preferred by users. It uses a model based recommendation approach, where prediction is modelled as a regression problem. Recommendations are usually fine-tuned to users. Hence predictions for one user pertain to that user only. The proposed collaborative filtering architecture is modelled in two phases. The first phase deals with identifying the current user's interests and forming their profile, finding users similar to the current profile and identifying the item vectors pertaining to similar users. Most recommendation systems stop at this level to provide recommendations. The proposed approach moves further by building a training matrix from the item vectors that is passed to the next phase. The second phase uses a boosted tree based ensemble to create a prediction model that is used for the final predictions. Experiments and comparisons indicate the

high effectiveness of the proposed model as it exhibits a considerable reduction in the Mean Absolute Error (MAE) and the Root-Mean-Square Error (RMSE) in comparison to existing state-of-the-art models.

The remainder of the paper is organized as follows: section 2 presents the literature review, section 3 presents the problem formulation, section 4 presents a detailed description of the proposed ensemble model, section 5 presents the experimental results and section 6 concludes the work.

2 Literature review

This section discusses some of the recent contributions in the domain of recommendation systems.

An artificial neural network based recommendation system that uses content-based modelling for predictions was proposed by Paradarami et al. in [11]. This work performs model based predictions by utilizing user reviews. Model based predictions are hybridized with ANN to perform enhanced predictions. A similar hybridized recommendation model specifically for e-learning environments was proposed by Chen et al. in [12]. Several hybrid versions of recommenders are currently on raise, like user specific hybrid recommender for offline optimization by Dooms et al. [13], an augmented matrix based hybrid system by Wu et al. [14], a latent factor based recommendation system by Zheng et al. [15] and a linear regression based collaborative recommendation model by Ge et al. [30].

Utilizing metaheuristics for recommendations have currently been on the raise due to the increase in data volume. A cuckoo search based collaborative filtering model was proposed by Katarya et al. in [16]. This model uses k-means clustering for user grouping and cuckoo search for the process of prediction. Other metaheuristic or evolutionary algorithm based recommendation systems include memetic algorithm and genetic algorithm based recommender system by Banati et al. [17], PSO based recommender system [18] and fuzzy ant based recommenders by Nadi et al. [19].

A weighting strategy based recommender that performs genre based clustering was proposed by Fremal et al. [20]. This method analyzes twelve weighting strategies in terms of MAE and RMSE to obtain the best weighting model for effective recommendations. A similar multiple clustering based recommendation model was proposed by Ma et al. in [21]. A trust and similarity based recommender for leveraging multiviews was proposed by Guo et al. in [22]. A prediction system to recommend complimentary products was proposed by McAuley et al. in [23]. This model concentrates in identifying substitutable versions of customers' interests. A coordinate based recommendation system SCoR was proposed by Papadakis et al. [32]. This model uses a combination of matrix factorization and collaborative filtering to improve the prediction process. A user perception based model was proposed by Chen et al. [33]. This is a critiquing based model that considers user's perception of products for the prediction process.

Although several models for recommendations are available, most of them follow the regular filtering mechanisms, resulting in huge data for processing, thereby increasing the computational complexity to a large extent.

3 Problem formulation

The collaborative filtering model has been formulated as a prediction problem, where the proposed Tree Based Ensemble (TBE) model predicts the probable ratings that will be provided by the user for a particular item.

Let CL be the set of customers, where $CL=\{C_1,C_2,C_3 \ldots C_m\}$ and P_C be the purchase list of a customer, where the item purchased i_x and corresponding rating given to the item r_x are the mandatory components. All available n items are contained in the items list $I=\{i_1,i_2,i_3 \ldots i_n\}$. The ratings are formulated as real numbers R in the interval $[r_{min}, r_{max}]$, where r_{min} and r_{max} are defined by the domain.

The problem is to predict a set of items from I for a customer C such that the customer would have a high probability of purchasing it, pertaining to constraints given in eq 1.

$$Prediction_C = \{i_x | i_x \in I \wedge i_x \notin P_C\} \quad (1)$$

4 Collaborative filtering using tree based boosted ensemble

Collaborative filtering is the process of predicting a user's interests based on their past behaviors. This paper proposes a model based collaborative filtering approach using a boosted ensemble. Algorithm for the proposed collaborative filtering architecture is given below.

Algorithm:

1. *Input user (C) for recommendation*
2. *Generate item list I_C from the customer purchase history*
3. *Generate rating list IR_C using Eq. 3*
4. *Identify similar users (NC) by selecting users with common item list using Eq. 4*
5. *Generate rating list by adding the item and ratings given by C and NC*
6. *Identify correlation between the rating vectors of C and NC using Eq. 6*
7. *Filter items with correlation levels higher than the similarity threshold ρ_{Thresh}*
8. *Identify categories of the selected items*
9. *Normalize categories to obtain the training matrix (T)*
10. *Create the recommender model by applying T to the boosted tree based ensemble*
11. *Predict ratings for all the available items*
12. *Filter n best items with highest ratings as recommendations to the user C*

The proposed collaborative filtering architecture has been modelled in two major phases namely; profile induced item matrix creation using feature augmentation and ensemble based predictions. The first phase collects,

filters and integrates data corresponding to the user for whom the recommendation is to be made. The second phase creates a boosted ensemble, trains it using the created training data and provides predictions.

4.1 Profile induced item matrix creation using feature augmentation

Recommendation systems are built for heterogeneous users. Every user's requirements is distinct, however there also exists slight similarities with other users in the system [24]. Hence it is important to identify and build appropriate profile for the current user for the model to be trained upon. Effectiveness of predictions depends entirely on the quality of the training data built at this phase. The process of building the user's profile is performed using Feature Augmentation. Feature Augmentation is the process of computing a set of features to be passed to the subsequent phase for evaluation.

The initial phase of this process is to generate an item list from the purchase history of the customer under analysis. This is given by

$$I_C = \{i_x | i_x \in I \wedge i_x \subset P_C\} \quad (2)$$

Where i_x is an item from the set of all items purchased by the customer.

The ratings pertaining to the item list I_C are integrated to obtain the user's preferences.

$$IR_C = \{(i_x, r_x) | (i_x, r_x) \subset P_C\} \quad (3)$$

Where r_x is the rating corresponding to item i_x.

The mere factor that the customer has purchased an item does not guarantee the person's affinity towards the product. Hence affinity levels for the product are obtained by integrating the ratings.

The next step is to identify users with interests similar to the current user *C*. Identifying similarities begins by identifying the commonalities existing between *C*, the current user, paired with all the other existing users (NC). This is given by

$$I_{Common} = I_C \cap I_{NC} \quad \forall \quad 1 \leq NC \leq m \wedge NC \neq C \quad (4)$$

Where I_C and I_{NC} correspond to items purchased by customers *C* and *NC*.

The common items identified in I_{Common} are integrated with their corresponding ratings from *C* and *NC*, to create the ratings matrix, which is given by,

$$IR_{Common} = \{(i_x, r_c, r_{nc}) | (i_x, r_c) \subset P_C \wedge (i_x, r_{nc}) \subset P_{NC}\} \quad (5)$$

Where r_c is the rating given to product i_x by customer *C* and r_{nc} is the rating given to product i_x by customer *NC*.

Correlation of ratings r_c and r_{nc} between the current customer *C* and every other customer *NC* is determined to identify the similarity levels between the two customers. Similarities are identified between two rating vectors, and a similarity identification model is used for the process [25]. Some of the common similarity measures are Euclidean distance, Minkowski distance and Pearson correlation [2]. Distance based measures requires the input vectors to be standardized prior to operations, while major advantage of a correlation based model is that they operate based on cosine similarities, hence do not require standardized values. This avoids the additional overhead of standardizing the input data. Hence the proposed TBE model uses Pearson correlation as the similarity measure identifier. TBE model uses all the items identified as common (entire population) to obtain the similarity, which is given by

$$\rho_{C,NC} = \frac{cov(R_C, R_{NC})}{\sigma_{R_C} \sigma_{R_{NC}}} \quad (6)$$

Where R_C and R_{NC} are the rating vectors corresponding to *C* and *NC* , the numerator calculates the covariance of R_C and R_{NC}, the denominator calculates the product of standard deviations of R_C and R_{NC}.

The final item set is obtained by filtering item data satisfying the similarity threshold (ρ_{Thresh}). The similarity threshold is domain and data dependent. The proposed TBE model sets a similarity threshold of 0.5 for analysis. Items corresponding to users with satisfied thresholds are considered for building the training matrix. The selection criteria for items is given by

$$I_{Selected} = \{i_x | i_x \in IR_{Common} \wedge i_x \in I_{NC} \wedge \rho_{C,NC} < \rho_{Thresh}\} \quad (7)$$

Item categorization plays a vital role in determining the details pertaining to items. Training data for TBE is constructed with the item categorization details, rather than the actual items. Broader and highly specific categorizations tend to provide more accurate results. The proposed model also deals with integrating items falling under multiple categories. Categories pertaining to items under $I_{Selected}$ are obtained. Item categorizations tend to be nominal rather than numeric. Hence they are normalized with 1-of-n encoding to obtain the numeric training data matrix (*T*) for training the ensemble model.

4.2 Ensemble based Predictions

Model based recommenders utilize a machine learning model to predict recommendations for a user. The proposed TBE model builds a boosted tree ensemble for prediction.

Ensemble modelling [26] is the process of incorporating multiple models for prediction, rather than relying on the results of a single model. Boosting is a machine learning ensemble aimed to reduce bias and variance in the prediction system to provide an effective prediction model. It is a supervised learning approach operating by creating a set of weak learners to form a single strong learner. This work uses decision trees to build the model for recommendations based on the training data [27].

The proposed boosting model operates by iteratively training the algorithm based on the resultant errors from previous iterations.

Let *DT(x)* be the base decision tree used for training. The process of prediction is given by

$$y' = DT(x) \quad (8)$$

Where y' is the prediction given by the decision tree model *DT*. However, being a weak learner, the predictions by *DT* will constitute errors *e*, which can be given by

$$e = y' - y \quad (9)$$

Where y is the actual solution and y' is the predicted solution.

The next level prediction model is built by integrating the error component e into the prediction model. This is given by

$$y'' = DT(x) + e \quad (10)$$

Similarly, the next level error is given by

$$e' = y - y'' \quad (11)$$

The next model training incorporates e' into the training process. This is iteratively performed until the error e reaches an acceptable threshold.

Training data for the recommendation problem is modelled with category based training matrix constructed from item ratings obtained from similar users. Rating corresponding to the item vector is incorporated as the class label for the training matrix (*T*). The training matrix is passed to the boosted decision tree and the trained model is obtained. The process of training matrix creation and prediction is repeated for each user individually on every recommendation requirement, to obtain result pertaining to an individual user.

Test data is obtained by considering the items not contained in the purchase list of customer *C*. The formulation of test data is given by

$$I_{Test} = \{i_x | i_x \in I \wedge i_x \notin P_C\} \quad (12)$$

Categories corresponding to I_{Test} are obtained and 1-of-n encoding is applied to obtain the test matrix. TBE is formulated as a regression model. Hence an error level of 0.001 is set as the acceptable error limit for the TBE model.

The results provide probable user ratings for each item. Recommendations can be provided by sorting the results in decreasing order and providing the top *n* rated products as probable recommendations.

5 Experimental results

The proposed TBE model is implemented using Python and uses MovieLens data [28, 29] for analysis. MovieLens is a benchmark dataset used to validate recommendation systems. The dataset pertaining to 1Million reviews is considered for evaluation. It contains details pertaining to 6040 users and provides reviews for 3952 movies. Ratings are provided on a 5 point scale. The dataset also contains categorizations of movies in terms of genres. A single movie sometimes belongs to multiple genres, providing scope for multiple options. The proposed model operates by considering movies as items and genres as the categorization parameters.

Recommendation models are usually measured in terms of Root-Mean-Square Error (RMSE) and Mean Absolute Error [13, 31].

$$MAE = \frac{1}{N}\sum_{i=1}^{n} |y_i - y'_i| \quad (13)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n} (y_i - y'_i)^2} \quad (14)$$

Where y_i and y_i' are the actual and the predicted ratings for the *N* test reviews. MAE measures the effectiveness of the predictions. Smaller MAE values exhibit better predictions. RMSE depicts the stability of the predictions, in other words, the prediction variance. Low MAE values represent a good predictor, while high RMSE values indicate high variability in predictions.

Performance of the proposed model is measured in terms of RMSE and MAE. Scalability of the model is measured by sampling the data from 100K reviews, moving up to 1 Million reviews. Exhibited results were attained by performing 1000 iterations and identifying the mean of the obtained predictions.

Mean Absolute Error (MAE) corresponding to datasets of various sizes is shown in figure 1. It could be observed that the proposed TBE model exhibits similar MSE values exhibiting low fluctuations irrespective of the data size. The best MAE was observed to be 0.51, and average MAE was observed to be 0.64, with fluctuation levels of ±0.1. This exhibits the stability of the proposed model irrespective of the data size.

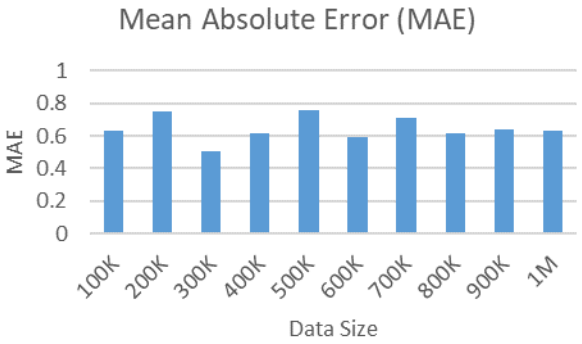


Figure 1: Mean Absolute Error Analysis for Varied Sized Input Data.

Root-Mean-Square Error (RMSE) corresponding to datasets of various sizes is shown in figure 2. It could be observed that the proposed TBE model exhibits similar RMSE values exhibiting low fluctuations irrespective of

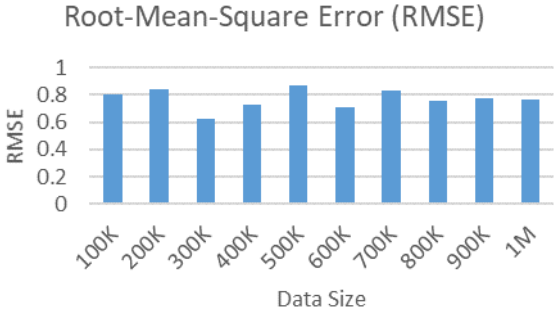


Figure 2: Root-Mean-Square Error Analysis for Varied Sized Input Data.

the data size. The best RMSE was observed to be 0.62, and average RMSE was observed to be 0.77, with fluctuation levels of ±0.1. This exhibits the low variability in prediction levels of TBE.

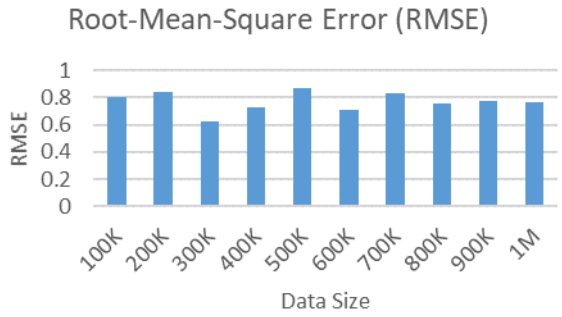


Figure 2: Root-Mean-Square Error Analysis for Varied Sized Input Data.

Enhanced performance of the proposed model is attributed to the two major factors, feature augmentation and the tree based boosted ensemble. Feature augmentation is based on the user's profile. Hence the input data contains several attributes depicting the user's profile. This results in the model being highly fine-tuned towards the user's requirements. This enables better predictions. further, usage of the boosting model ensures that every wrong prediction increases the weight of the instance. This enables even rarely found instances to have a significant impact on the final prediction process. These factors enable enhanced results in the proposed model.

The actual performance values for TBE is tabulated and shown in table 1. Best performances are shown in bold. Moderate MAE and RMSE values indicate effectively reduced error levels and low variability levels in predictions. Time requirements for the proposed model exhibits linear time requirements by the TBE model.

Dataset	MAE	RMSE	Time(s)
100K	0.630135	0.806763	9.197
200K	0.746131	0.841447	20.662
300K	**0.506511**	**0.625448**	51.306
400K	0.617964	0.724338	88.696
500K	0.757933	0.871471	94.607
600K	0.590549	0.705942	112.125
700K	0.710233	0.835888	105.766
800K	0.618823	0.753463	108.144
900K	0.641097	0.77633	113.657
1M	0.629914	0.763647	118.935

Table 1: Performance Levels on Data with Varied Sizes.

Comparison of the proposed model is performed with the weighted strategy based model (SW I, MLR and CM II) proposed by Fremal et al. [20] and K-Means and Cuckoo Search based model proposed by Katarya et al. [16]. Both these models are recent and also considers the MovieLens data for their prediction process. Hence this work considers the two models for comparison.

A comparison on MAE values of the proposed model with SW I, MLR, CM II and K-Means Cuckoo is shown in figure 3. It could be observed that the proposed TBE model exhibits lowest MAE levels, exhibiting reduction levels in the range of 6% to 14%.

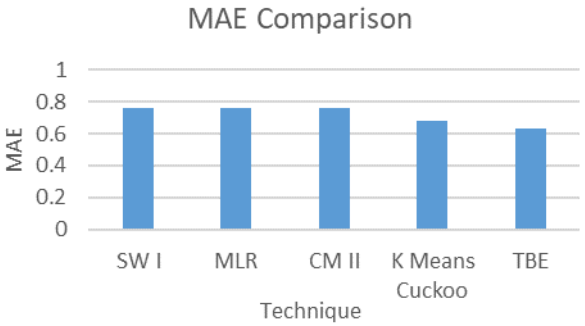


Figure 3: Mean Absolute Error Comparison.

A comparison on RMSE values of the proposed model with SW I, MLR, CM II and K-Means Cuckoo is shown in figure 4. It could be observed that the proposed TBE model exhibits lowest RMSE levels (0.76), while the other models exhibits RMSE levels >0.9. This exhibits that the variance levels of the proposed model is low compared to the other models. Low variability levels indicate high prediction reliability of the proposed TBE model.

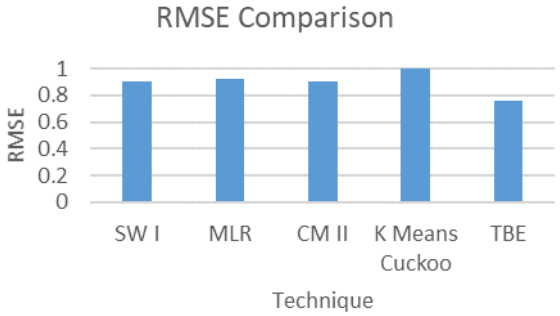


Figure 4: Root-Mean-Square Error Comparison.

6 Conclusion

Recommendations have become one of the major requirements in the current information rich world. However, the voluminous data available for the recommendation engines poses a huge challenge. This paper proposes a feature augmentation based tree bagging ensemble model, TBE for recommendations. TBE, being an iterative model uses a weak classifier, hence the computational complexities pertaining to TBE was observed to be very low. Further, the repeated data filtering process in the first phase reduces data to a large extent. This further reduces the computational complexities, hence speeding up the prediction process to a considerable extent. This aids in handling large datasets effectively, indicating enhanced scalability levels.

The major advantage of TBE is that it uses the available data for the current user and user's similar to the current user. Hence TBE has the capability to identify

even hidden patterns from the available data. Further, this process of prediction solves the data sparsity issue that affects collaborative filtering approaches. Limitations of the proposed model are that cold start problem has not been handled. Future extensions of the proposed model will be designed by incorporating user's demographic data, which can enable solving the cold start issue.

References

[1] Vickery, A., and Vickery, B. C. (2005). *Information science in theory and practice*. De Gruyter. https://doi.org/10.1515/9783598440083

[2] Ricci, F., Rokach, L., and Shapira, B. (2011). *Introduction to recommender systems handbook* Boston, MA: Springer US . pp. 1-5. https://doi.org/10.1007/978-0-387-85820-3_1

[3] Karlgren, Jussi. "A digital bookshelf: original work on recommender systems"

[4] Lohr, S. (2009). A $1 million research bargain for Netflix, and maybe a model for others. *The New York Times*, *22*.

[5] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17* (6), 734–749. https://doi.org/10.1109/TKDE.2005.99

[6] Prem Melville and Vikas Sindhwani, (2010) Recommender Systems, *Encyclopedia of Machine Learning*, pp. 1-9.

[7] Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 253-260. https://doi.org/10.1145/564376.564421

[8] Sarwar, B. M., Karypis, G., Konstan, J., & Riedl, J. (2002). Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *Proceedings of the fifth international conference on computer and information technology*, Vol. 1.

[9] Mooney, R.J., and Roy, L. (1999). Content-based book recommendation using learning for text categorization. *In Workshop Recom. Sys.: Algo. and Evaluation.* https://doi.org/10.1145/336597.336662

[10] Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: an introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9780511763113

[11] Paradarami, T.K., Bastian, N.D. and Wightman, J.L., (2017). A hybrid recommender system using artificial neural networks. *Expert Systems with Applications*, vol. 83, pp.300-313. https://doi.org/10.1016/j.eswa.2017.04.046

[12] Chen, W., Niu, Z., Zhao, X., and Li, Y. (2014). A hybrid recommendation algorithm adapted in e-learning environments. *World Wide Web*, doi: 10. 1007/s11280-012-0187-z .vol. 17(2), pp.271-284. https://doi.org/10.1007/s11280-012-0187-z

[13] Dooms, S., Pessemier, T., and Martens, L. (2015). Offline optimization for user-specific hybrid recommender systems. *Multimedia Tools and Applications*, vol. 74(9) , pp.3053-3076. https://doi.org/10.1007/s11042-013-1768-2

[14] Wu, M.L., Chang, C.H., and Liu, R.Z. (2014). Integrating content-based filtering with collaborative filtering using co-clustering with augmented matrices. *Expert Systems with Applications*, vol. 41(6),pp.2754-2761. https://doi.org/10.1016/j.eswa.2013.10.008

[15] Zheng, Y. (2014). Semi-supervised context-aware matrix factorization: Using contexts in a way of "latent" factors. *In Proceedings of the 29th annual ACM symposium on applied computing* . In SAC New York, NY, USA: ACM .vol.14, pp.292-293. https://doi.org/10.1145/2554850.2555195

[16] Katarya, R. and Verma, O.P., (2016). An effective collaborative movie recommender system with cuckoo search. *Egyptian Informatics Journal*. https://doi.org/10.1016/j.eij.2016.10.002

[17] Banati, H. and Mehta, S., (2010). A Multi-Perspective Evaluation of ma and ga for collaborative Filtering Recommender System. *International journal of computer science & information Technology (IJCSIT)*, vol. *2*(5), pp.103-122. https://doi.org/10.5121/ijcsit.2010.2508

[18] Alam, S., Dobbie, G. and Riddle, P., (2011), May. Towards recommender system using particle swarm optimization based web usage clustering. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* . Springer, Berlin, Heidelberg.pp.316-326. https://doi.org/10.1007/978-3-642-28320-8_27

[19] Nadi, S., Saraee, M., Bagheri, A. and Davarpanh Jazi, M., (2011). FARS: Fuzzy ant based recommender system for web users. *International Journal of Computer Science Issues*, vol.8 (1), pp.203-209. https://doi.org/10.20533/ijmip.2042 .4647.2011.0001

[20] Frémal, S. and Lecron, F., (2017). Weighting strategies for a recommender system using item clustering based on genres. *Expert Systems with Applications*, vol.77, pp.105-113. https://doi.org/10.1016/j.eswa.2017.01.031

[21] Ma, X., Lu, H., Gan, Z., and Zhao, Q. (2016). An exploration of improving prediction accuracy by constructing a multi-type clustering based recommendation framework. *Neurocomputing*, vol.191, pp. 388–397. https://doi.org/10.1016/j.neucom.2016.01.040

[22] Guo, G., Zhang, J., and Yorke-Smith, N. (2015). Leveraging multiviews of trust and similarity to enhance clustering-based recommender systems. *Knowledge-Based Systems*, vol. 74 , pp.14–27. https://doi.org/10.1016/j.knosys.2014.10.016

[23] McAuley, J., Pandey, R., and Leskovec, J. (2015).a Inferring networks of substitutable and complementary products. *In Proceedings of the*

twenty first ACM SIGKDD international conference on knowledge discovery and data mining . In KDD15 New York, NY, USA: ACM. pp.785-794. https://doi.org/10.1145/2783258.2783381

[24] Kunaver, M., & Požrl, T. (2017). Diversity in recommender systems–A survey. *Knowledge-Based Systems*, Vol. 123, pp. 154-162. https://doi.org/10.1016/j.knosys.2017.02.009

[25] Sorensen, S. (2012). Accuracy of similarity measures in recommender systems. In Proceedings of the 29th SemiAnnual Computer Science Senior Seminar Conference (Minnesota, USA.

[26] Rokach, L. (2010). "Ensemble-based classifiers". *Artificial Intelligence Review*. Vol. **33**, pp. 1–39. https://doi.org/10.1007/s10462-009-9124-7

[27] Breiman, L. (1997) "Arcing The Edge"

[28] https://grouplens.org/datasets/movielens/

[29] Harper, F.M., and Konstan, J.A. (2015). The MovieLens datasets: History and context. ACM *Transactions on Interactive Intelligent Systems*, doi: 10.1145/2827872 .vol. 5(4), pp.19:1-19:19. https://doi.org/10.1145/2827872

[30] Ge, X., Liu, J., Qi, Q., & Chen, Z. (2011, July). A new prediction approach based on linear regression for collaborative filtering. In *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* (Vol. 4, pp. 2586-2590). IEEE. https://doi.org/10.1109/FSKD.2011.6020007

[31] Ge, X., Liu, J., Qi, Q., & Chen, Z. (2011). A new prediction approach based on linear regression for collaborative filtering. IEEE *Eighth International Conference* In *Fuzzy Systems and Knowledge Discovery (FSKD)*, Vol. 4, pp. 2586-2590. https://doi.org/10.1109/FSKD.2011.6020007

[32] Papadakis, H., Panagiotakis, C., & Fragopoulou, P. (2017). SCoR: a synthetic coordinate based recommender system. *Expert Systems with Applications*, *79*, 8-19. https://doi.org/10.1016/j.eswa.2017.02.025

[33] Chen, L., Yan, D., & Wang, F. (2019). User perception of sentiment-integrated critiquing in recommender systems. *International Journal of Human-Computer Studies*, *121*, 4-20. https://doi.org/10.1016/j.ijhcs.2017.09.005

A Solution to the Problem of the Maximal Number of Symbols for Biomolecular Computer

Jacek Waldmajer
Institute of Computer Science, University of Opole, Oleska 48, 45-052 Opole, Poland
E-mail: jwaldmajer@uni.opole.pl

Sebastian Sakowski
Faculty of Mathematics and Computer Science, University of Lodz, Banacha 22, 90-238 Lodz, Poland
E-mail: sebastian.sakowski@wmii.uni.lodz.pl

Keywords: biomolecular computer, biomolecular systems, DNA computing

Received: March 15, 2019

The authors present a solution to the problem of generating the maximum possible number of symbols for a biomolecular computer using restriction enzyme $BbvI$ and ligase as the hardware, and transition molecules built of double-stranded DNA as the software. The presented solution offers an answer to the open question, in the algorithm form, of the maximal number of symbols for a biomolecular computer that makes use of the restriction enzyme $BbvI$.

Povzetek: Razvit je nov način izračuna največjega števila simbolov za biomulekularni računalnik.

1 Introduction

The beginnings of research into possibilities of applying biomolecules to control biological systems, and also to construct computers, are to be found in theoretical works of the 1960s (Feynman 1961). Then, in the 1980s, Charles Bennett (1982, Bennett and Landauer 1985) pointed to potential possibilities of application of biomolecules to construct energy-efficient nanodevices. However, the world had to wait to see the first practical experiments realizing simple calculations with the use of biochemical reactions until the mid-1990s, when Leonard Adleman (1994) solved the problem of Hamilton's path in graph, using exclusively a biomolecule for this purpose. Successive research revealed the possibility of spontaneous formation of multidimensional structures built from biomolecules, which were made with the use of the conception of self-assembly (Whitesides et al. 1991, Seeman 2001, Gopinath et al. 2016). The multidimensional DNA structures made it possible to realize fractals, e.g., ones of Sierpiński triangle type (Rothemund 2004), which revealed a great potential in calculations based on self-assembly. In 2006, Paul Rothemund (2006) made use of self-assembling DNA molecules to obtain different multidimensional biomolecular structures. Properly prepared DNA molecules also made it possible to carry out a theoretical simulation of Turing machine (Rothemund 1995). Prior to this, in 2001 (Benenson et al. 2001) a practically acting non-deterministic finite automaton based on such DNA molecules, restriction enzyme FokI and DNA ligase was presented. In successive research, it was proved experimentally that such an automaton can work without the use of ligase enzyme (Benenson et al. 2003, Chen et al. 2007) and its complexities were extended in practical experiments, ones understood as the number of states using numerous restriction enzymes (Sakowski et al. 2017). It is worth adding that it was with success that laboratory experiments were carried out, in which this biomolecular system was applied to medical diagnosis and treatment (Benenson et al. 2004) and also to simple logical inference (Ran 2009). In another work which dealt with possibilities of applying DNA molecules, a challenge was taken up to not only increase the number of states of such an automaton (Unold et al. 2004), but also that of symbols possible for an automaton built from DNA (Soreni et al. 2005). Moreover, presented the notion of biomolecular automaton, informally characterized in the papers of Rothemund (1995), Benenson et al. (2001), Soreni et al (2005), was presented in a formal way (as a mathematical model called a tailor automaton in a new theory of tailor automata) in the paper Waldmajer et al. (2019).

In the above-mentioned work, Soreni and co-workers (Soreni et al. 2005) put forward a 3-state 3-symbol biomolecular automaton which used the restriction enzyme BbvI as well as considered the problem of determining the maximal number of symbols for the constructed biomolecular automaton. On the basis of the conducted assessment they pointed out that it is possible to construct 40 symbols, each of which is composed of 6 pairs of nucleotides. However, in their work, they pointed to merely 37 such symbols, including one which was erroneously determined. Consequently, they opened the following issue (p. 3937): *It is still an open question whether the maximal number of 6-bp sequences that produce distinct 4-bp sticky ends in both*

strands is 40. It is with reference to this open question that the authors of the present work undertook and managed to solve the problem mentioned by Soreni et al. in their work (2005) through: (1) indicating 40 symbols (see Tab. 4) which make the solution to the open problem, (2) proposing the idea of working of an algorithm that enables to generate 40 symbols for a biomolecular automaton using the restriction enzyme BbvI, and (3) formulating two general problems in the sphere of generating symbols for biomolecular automata which use one restriction enzyme (among which a biomolecular automaton using the restriction enzyme BbvI is a particular case) and more than one restriction enzyme.

The second section presents the idea of constructing and working of a 3-state 3-symbol biomolecular automaton using the restriction enzyme BbvI as presented by Soreni and co-workers in their work (Soreni et al. 2005). In the third section the conception of working of an algorithm generating the maximal number of symbols for a biomolecular automaton using the restriction enzyme BbvI was presented together with a discussion of various undesired situations which may occur in the course of working of a biomolecular automaton that makes use of one restriction enzyme (in particular for the restriction enzyme BbvI). In the last section, there were formulated two general problems of generating the maximal number of symbols for a certain class of biomolecular automata using one or more than one restriction enzymes.

2 Biomolecular finite automaton and the idea of its actions

In this section, we make a presentation of the 3-state 3-symbol biomolecular finite DNA automaton (see Fig. 1), which was presented by Soreni and co-workers (Soreni et al. 2005). The automaton uses the restriction enzyme BbvI, ligase enzyme and DNA double-stranded fragments (input molecule, set of transition molecules and set of detection molecules). The double-stranded DNA fragments include the adenine, cytosine, guanine, and thymine bases marked as A, C, G and T, respectively.

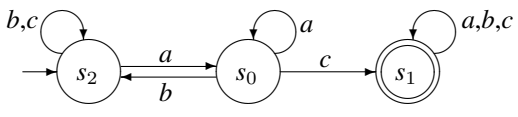


Figure 1: Graph representing a 3-state 3-symbol deterministic finite automaton M_1.

The task of the BbvI restriction enzyme is to cut the double-stranded DNA after recognizing a specific sequence (see Fig. 2A) in the double-stranded DNA.

The BbvI restriction enzyme will cut the double-stranded DNA after the 8th nucleotide in the DNA strand in the 5′-3′ direction and after the 12th nucleotide in the DNA strand in the 3′-5′ direction from the recognized specific sequence (see Fig. 2B).

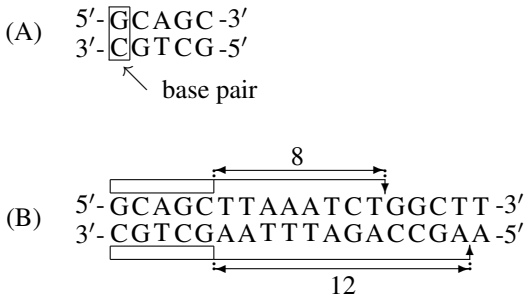


Figure 2: (A) Specific sequence recognized by the BbvI restriction enzyme. (B) The action of restriction endonuclease: BbvI.

The task of the ligase enzyme is to ligate the two double-stranded DNAs having complementary sticky ends (see Fig. 4A and 4B), where a sticky end is a single-stranded DNA at the end of a double-stranded DNA. In the given sense, the sticky end 'TTTA' of a single-stranded DNA (see Fig. 4A) is complementary to a sticky end 'AAAT' of the other double-stranded DNA (see Fig. 4B). The result of their ligation is one double-stranded DNA (see Fig. 4C).

Both the restriction enzyme BbvI and the ligase enzyme play the key role in the action of a biomolecular automaton, determining, respectively: the operation of cutting of a fragment of the double-stranded DNA and the operation of ligating of two fragments of double-stranded DNAs.

The input molecule (see Fig. 3) is a double-stranded DNA fragment in which it is possible to distinguish the following three basic parts: the input word x consisting of the symbols a, b and c (x = acb), the terminal symbol and the base sequence. At the both ends of the input molecule there occur additional base pairs and their occurrence is determined by the properties related to the action of the restriction enzyme.

To construct an input word of the 3-state 3-symbol deterministic finite automaton, the following three symbols: a, b and c (see Fig. 5) were used. These symbols were coded by means of six base pairs. Besides the aforementioned symbols, the additional terminal symbol t was introduced. This symbol is coded by means of the same number of base pairs as the symbols a, b and c. This symbol was used to acquire an output molecule which is used to determine whether the automaton has finished acting in the required state and has accepted the input word x.

The base sequence consists of a certain number of base pairs, contains a specific sequence recognizable by the BbvI restriction enzyme, and makes it possible to define the start state by determining the cut place of the input molecule by the BbvI restriction enzyme (cf. Fig. 3 and Fig. 2B). Let us note that the term "base sequence" did not appear in work Soreni et al. (2005). Introducing this term is meant to clearly determine the manner of setting the start state of a biomolecular automaton. According to the idea contained in the work of Soreni and co-workers

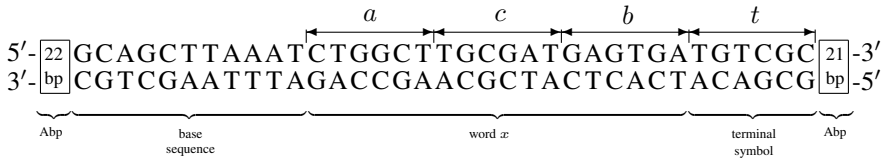


Figure 3: Input molecule containing the input word $x = acb$; Abp – Additional base pairs.

Table 1: Connection of the states s_0, s_1 and s_2 of a biomolecular automaton with the permanent cut places of the symbols a, b, c and t of the biomolecular automaton.

state	*symbol a*	*symbol b*	*symbol c*	*symbol t*
s_0	5′-|CTGGCT-3′ 3′-GACC|GA-5′	5′-|GAGTGA-3′ 3′-CTCA|CT-5′	5′-|TGCGAT-3′ 3′-ACGC|TA-5′	5′-|TGTCGC-3′ 3′-ACAG|CG-5′
s_1	5′-C|TGGCT-3′ 3′-GACCG|A-5′	5′-G|AGTGA-3′ 3′-CTCAC|T-5′	5′-T|GCGAT-3′ 3′-ACGCT|A-5′	5′-T|GTCGC-3′ 3′-ACAGC|G-5′
s_2	5′-CT|GGCT-3′ 3′-GACCGA|-5′	5′-GA|GTGA-3′ 3′-CTCACT|-5′	5′-TG|CGAT-3′ 3′-ACGCTA|-5′	5′-TG|TCGC-3′ 3′-ACAGCG|-5′

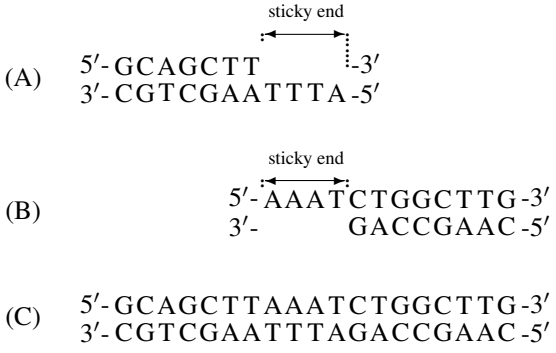


Figure 4: (A) and (B) Double-stranded DNAs with the complementary sticky ends. (C) The result of a ligation between the double-stranded DNAs with the complementary sticky ends included in (A) and (B).

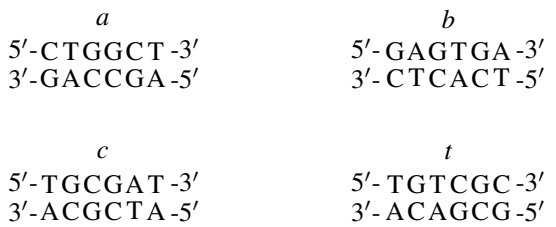


Figure 5: Symbols a, b, c and the terminal symbol t.

(Soreni et al. 2005), the reading of a symbol in a certain state of the automaton is identified with the cutting of the double-stranded DNA by the BbvI restriction enzyme in the area of a symbol, in a determined (permanent) place of the DNA strand, in the 5′-3′ direction and in a determined (permanent) place of the DNA strand in the 3′-5′ direction. Tab. 1 presents a connection between the states and two permanent cut places of the symbols.

In accordance with the input molecule presented on Fig. 3, the first cutting input molecule with the use of the restriction enzyme BbvI will follow in the area of the symbol a, which corresponds to the state s_2. In this way, in the state s_2, the symbol a was read and a fragment of DNA was obtained as presented on Fig. 6. In this sense, the state s_2 is a initial state. Adding to the base sequence of one or two pairs of nucleotides can set the start state to be the following: s_1 or s_0, respectively.

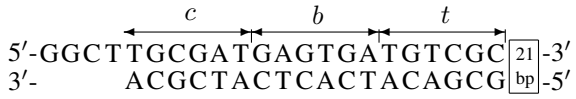


Figure 6: Double-stranded DNA fragment obtained after BbvI acting on the input molecule.

The set of transition molecules is used to implement a set of transitions in the 3-state 3-symbol deterministic finite automaton. We obtain transition from one state to the other (the same or another state), upon reading a symbol, through ligating with the use of ligase enzyme, of a DNA fragment obtained on Fig. 6 with one of the transition molecules. Each transition molecule contains a specific sequence recognizable by the BbvI restriction enzyme and the additional base pairs. Exemplary transition molecules are presented in Fig. 7: the transition molecule presented on Fig. 7A enables transition from the state s_0 to that of s_1 after reading the symbol c; the transition molecule presented on Fig. 7B enables transition from the state s_1 to that of s_1 after reading the symbol b; the transition molecule presented on Fig. 7C enables transition from the state s_2 to that of s_0 after reading the symbol a.

For each state, one detection molecule is constructed and thus a set of detection molecules is specified (see Fig. 8). It should be noted that the detection molecules have different numbers of additional base pairs, which makes it possible to determine laboratorily the state in which the automaton finished its action.

The beginning of the work: the BbvI, ligase enzyme, many copies of the transition molecules and many copies

(A) 5′-[35 bp]GCAGCT -3′
3′-CGTCGAACGC-5′

(B) 5′-[39 bp]GCAGCTT -3′
3′-CGTCGAATCAC-5′

(C) 5′-[35 bp]GCAGCTATT -3′
3′-CGTCGATAACCGA-5′

Figure 7: Selected transition molecules used in the transition function: (A) T_1: $(s_0, c) \rightarrow s_1$, (B) T_2: $(s_1, b) \rightarrow s_1$, (C) T_3: $(s_2, a) \rightarrow s_0$.

(A) 5′-[54 bp]ACAG-5′ -3′ (B) 5′-[200 bp]CAGC-5′ -3′

(C) 5′-[300 bp]AGCG-5′ -3′

Figure 8: (A) Detection molecule D_1 for the state s_0. (B) Detection molecule D_2 for the state s_1. (C) Detection molecule D_3 for state s_2.

of the detection molecules are placed in a laboratory tube; the final addition is many copies of the input molecule. After these elements have been mixed in the test tube, the biomolecular automaton starts its action. In successive steps there follows reading of the symbol a in the state s_2 (see Fig. 9a), making use of the transition molecule shown in Fig. 7C to transition from the state s_2 to that of s_0 after reading the symbol a (see Fig. 9b), reading of the symbol c in the state s_0 (see Fig. 9c), using the transition molecule presented in Fig. 7A to transition from the state s_0 to the state s_1 after reading the symbol c (see Fig. 9d) reading the symbol b in the state s_1 (see Fig. 9e), using the transition molecule presented in Fig. 7B and reading the terminal symbol t in the state s_1 (see. Fig. 9f-g). In the last step there follows ligation of a fragment of double-stranded DNA presented in Fig. 9g with one of the detection molecules (see Fig. 8B). As a result of ligation of these DNA fragments an output molecule is formed (see Fig. 9h), which – from the laboratory point of view – serves to determine the end state of a biomolecular finite automaton.

3 Algorithm for the problem of the maximal number of symbols

3.1 The formal aparatus used in the description of the algorithm

Let the set $\Delta = \{\text{A, C, G, T}\}$ and the function σ, which is bijection of the set Δ on Δ, which is defined in the following way: σ(A) = T, σ(T) = A, σ(C) = G and σ(G) = C be given. The set Δ is called a *set of nucleotides*, the elements of the set Δ are called *nucleotides*, and the function σ is called *complementarity of nucleotides*.

We call any finite sequence of nucleotides of the set Δ as a *word*. The word x which is the sequence $X_1, X_2, \ldots, X_j$ of nucleotides of the set Δ ($X_i \in \Delta$, $0 < i \leq j \in N$) is written as follows $x = X_1X_2 \ldots X_j$. The number of the elements of the sequence x is called the *length of the word* x (denoted symbolically: $|x|$), while the i-th nucleotide of the word x (the i-th element of the word x) as $x(i)$. The set of all the words formed from the nucleotides of the set Δ, whose length is greater than zero, is denoted as Δ^+.

Let $\Delta^+ \ni x = X_1X_2 \ldots X_j$ ($X_i \in \Delta$, $0 < i \leq j \in N$) and $\Delta^+ \ni y = Y_1Y_2 \ldots Y_j$ ($Y_i \in \Delta$, $0 < i \leq j \in N$). We call the word $X_j \ldots X_2X_1$ an *opposite word* (we denote symbolically: x^{-1}) to the word x. We call the word $xy = X_1X_2 \ldots X_iY_1Y_2 \ldots Y_j$ a *concatenation* xy of two words x and y such that $\Delta^+ \ni x = X_1\ X_2 \ldots X_i$ ($X_i \in \Delta$, $0 < i \in N$) and $\Delta^+ \ni y = Y_1Y_2 \ldots Y_j$ ($Y_j \in \Delta$, $0 < j \in N$). We say that the word x is *included in the word y, beginning with the k-th* ($1 \leq k \in N$) *position* (we denote symbolically: $x \subseteq_k y$), if $k + |x| \leq |y| + 1$ and $\exists u, v \in \Delta^*(y = uxv \wedge |u| = k - 1)$. The word x is a *sub-word of y* (we denote symbolically: $x \subseteq y$) when the word x is included in the word y, beginning with a certain position k, i.e., $x \subseteq y \Leftrightarrow \exists k(x \subseteq_k y)$. The word x is a prefix of the word y, when $x \subseteq_1 y$. The word x is a suffix of the word y, when x^{-1} is the prefix of the word y^{-1}.

The introduced notion of complementarity of nucleotides and the introduced denotations make it possible to define the function which will be called *complementariness of words*. The mapping Ξ: $\Delta^+ \rightarrow \Delta^+$ defined in the following way: $\Xi(x) = y$, where $|y| = |x|$ and $y(i) = \sigma(x(i))$ for each $i \in \{1, \ldots, |y|\}$ and, for $x \in \Delta^+$ is called *complementarity of words*.

Let $\Delta^+ \ni x = X_1X_2X_3X_4$ ($X_i \in \Delta$, $0 < i \leq j \in N$) and $\Delta^+ \ni y = Y_1Y_2Y_3Y_4$ ($Y_i \in \Delta$, $0 < i \leq l \in N$). The words x and y are *synthesable over the length* 3, when there exists the word $u \in \Delta^+$ of the length 3 being the suffix of the word x and the prefix of the word y. The concatenation of the synthesable words x and y over the length 3 is the word $z = [x, y]_3$, where $z = X_1X_2X_3X_4Y_4$.

3.2 Description of the algorithm

The idea of the algorithm of generating the maximal number of symbols for a biomolecular automaton using the restriction enzyme BbvI will be characterized through four stages, which are distinguished in the algorithm: the initial stage, the stage of deployment and verification, the stage of generation and the final stage. At each of the indicated stages we make use only of strands of symbols in the direction 5′-3′, since having strands of symbols in the direction 5′-3′, we can – by means of the principle of complementarity of nucleotides – obtain strands of symbols in the direction 3′-5′.

Let the set $\mathcal{A}_0$ of all 4-element sequences of nucleotides be given: $\mathcal{A}_0 = \{x : x \in \Delta^+ \wedge |x| = 4\}$=$\{\text{AAAA, AAAC, AAAG, \ldots, TTTG, TTTT}\}$. At the initial stage, we remove words (4-element sequences of nucleotides): AATT, ACGT, AGCT, ATAT, CCGG, CATG, CTAG, CGCG, GGCC, GATC, GTAC, GCGC, TTAA,

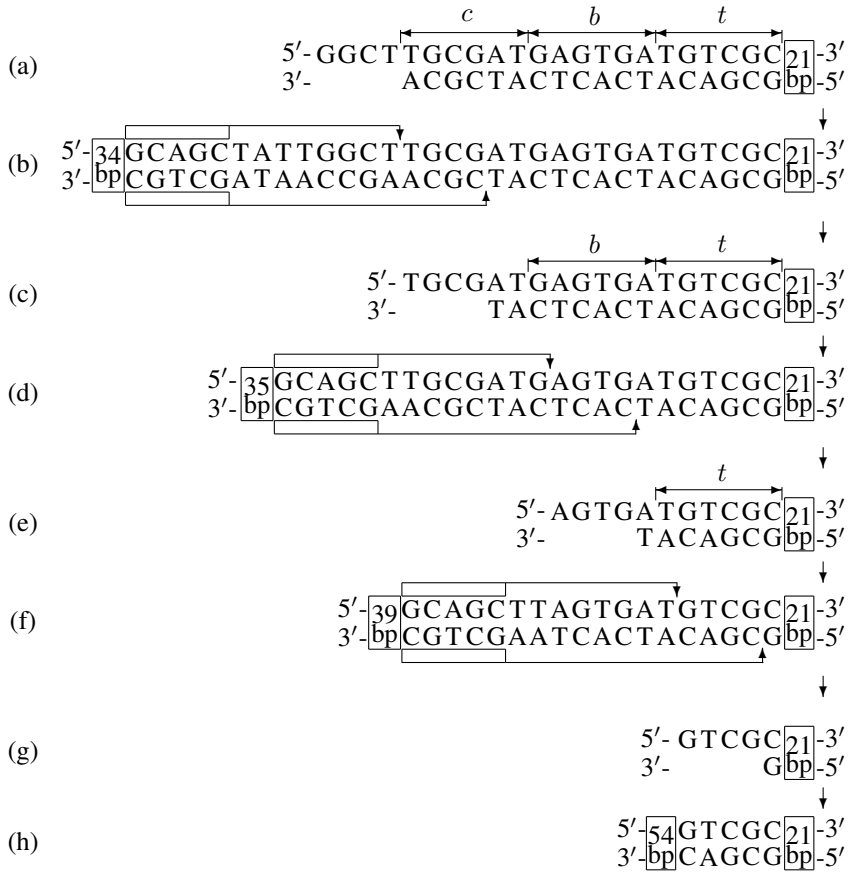


Figure 9: Control serving the reading of symbols of the word acb from the input molecule and obtaining an output molecule in the biomolecular automaton using the enzyme BbvI.

TCGA, TGCA, TATA from the set $\mathcal{A}_0$ of all 4-element sequences of nucleotides.

The appearance of the indicated sixteen words (4-element sequences of nucleotides) causes a biomolecular automaton to malfunction due to the possibility of ligation of a transition molecule with itself – each of the transition molecules exists in multi copies.

Let the transition molecule be given, in which we use the sticky end: CATG (see Fig. 10A). Let us note that this molecule occurs in many copies. Thus, as a result of action of the biomolecular automaton and ligation of one copy of the transition molecule T_{NS_1} (cf. Fig. 10A) with another copy of the same transition molecule there forms the double-stranded fragment of DNA presented in Fig. 10B. In consequence, this causes the number of copies of the molecule T_{NS_1}, to be limited, which can be made use of in further computations carried out by the biomolecular automaton.

So as to prevent the possibility of ligation of copies of the same transition molecule, it is necessary to remove from the set $\mathcal{A}_0$ the words which satisfy the following condition:

$$(*)\ \ x^{-1} = \Xi(x), \text{ where } x \in \mathcal{A}_0.$$

In this way, we reject sixteen words, given earlier, from the set $\mathcal{A}_0$ and as in consequence we obtain the set:

$$\mathcal{A}_1 = \{x : x \in \mathcal{A}_0 \wedge x^{-1} \neq \Xi(x)\} =$$

Figure 10: (A) Transition molecule T_{NS_1} using the sticky end: CATG. (B) Double-stranded fragment of DNA formed as a result of ligation of two copies of the transition molecule presented in (A).

$$\{x : x \in \Delta^+ \wedge |x| = 4 \wedge x^{-1} \neq \Xi(x)\},$$

where the number of the elements of the set $\mathcal{A}_1$ amounts to 240. Availing ourselves of the elements of the set $\mathcal{A}_1$, we form the maximal set $\mathcal{A}_2$ of pairs of the elements in the following manner:

$$\mathcal{A}_2 = \{(x, y) : x, y \in \mathcal{A}_1 \wedge x^{-1} = \Xi(y)\},$$

where the number of the elements of the set $\mathcal{A}_2$ amounts to 240. Then, using the set $\mathcal{A}_2$, we form the set $\mathcal{A}_3$ of pairs in the following way: the set $\mathcal{A}_3$ is the set $\mathcal{A}_2$, from which we removing certain pairs according to the principle of (P). The principle of (P): if the pairs (x, y) and (y, x) belong to the set $\mathcal{A}_2$, then we will remove from the set $\mathcal{A}_2$ a pair whose first element of the pair, comparing the both first el-

Table 2: Part I: 120 pairs (x,y) of four-element sequences of nucleotides.

No	x	y	No	x	y	No	x	y
1	AAAA	TTTT	21	ACCC	GGGT	41	AGTC	GACT
2	AAAC	GTTT	22	ACCG	CGGT	42	AGTG	CACT
3	AAAG	CTTT	23	ACCT	AGGT	43	ATAA	TTAT
4	AAAT	ATTT	24	ACGA	TCGT	44	ATAC	GTAT
5	AACA	TGTT	25	ACGC	GCGT	45	ATAG	CTAT
6	AACC	GGTT	26	ACGG	CCGT	46	ATCA	TGAT
7	AACG	CGTT	27	ACTA	TAGT	47	ATCC	GGAT
8	AACT	AGTT	28	ACTC	GAGT	48	ATCG	CGAT
9	AAGA	TCTT	29	ACTG	CAGT	49	ATGA	TCAT
10	AAGC	GCTT	30	AGAA	TTCT	50	ATGC	GCAT
11	AAGG	CCTT	31	AGAC	GTCT	51	ATGG	CCAT
12	AAGT	ACTT	32	AGAG	CTCT	52	ATTA	TAAT
13	AATA	TATT	33	AGAT	ATCT	53	ATTC	GAAT
14	AATC	GATT	34	AGCA	TGCT	54	ATTG	CAAT
15	AATG	CATT	35	AGCC	GGCT	55	CAAA	TTTG
16	ACAA	TTGT	36	AGCG	CGCT	56	CAAC	GTTG
17	ACAC	GTGT	37	AGGA	TCCT	57	CAAG	CTTG
18	ACAG	CTGT	38	AGGC	GCCT	58	CACA	TGTG
19	ACAT	ATGT	39	AGGG	CCCT	59	CACC	GGTG
20	ACCA	TGGT	40	AGTA	TACT	60	CACG	CGTG

ements of the pairs: (x, y) and (y, x), is lexicographically posterior (see Examp. 1). Tables Tab. 2 and Tab. 3 present 240 elements forming 120 pairs (x, y) of the set $\mathcal{A}_3$, where $x, y \in \mathcal{A}_1$.

Example 1: Let us note that the pairs (AAAA, TTTT), (TTTT, AAAA) $\in \mathcal{A}_2$. The first (element: AAAA) of the first pair (AAAA, TTTT) is lexicographically prior to the first element (element: TTTT) of the second pair (TTTT, AAAA). Thus, the pair (AAAA, TTTT) belongs to the set $\mathcal{A}_3$, and the pair (TTTT, AAAA) does not belong to $\mathcal{A}_3$.

Let us consider pair (x, y)=(AAAA, TTTT) from Tab. 2 (No 1) and two transition molecules in the biomolecular automaton with sticky ends: AAAA and TTTT (see Fig. 11A and Fig. 11B). As a result of ligation of these transition molecules is formed the double-stranded fragment of DNA presented in Fig. 11C. As a consequence, this causes the number of the copies of the molecules T_{NS_2} and T_{NS_3}, to be limited, which may be used in further calculations done by the biomolecular automaton. In connection with this, in the algorithm of generating the maximal number of symbols for a biomolecular automaton using the restriction enzyme BbvI only one element of each of the given 120 pairs of the set $\mathcal{A}_3$ should be used. Selecting individual elements of the successive pairs in this manner, we obtain the family $\mathcal{P}(A_1)$ of maximal sets $B \subset A_1$, such that for each $x, y \in B$ the condition holds,

$$(**)\ x^{-1} \neq \Xi(y).$$

The indicated condition (**) prevents the formation of transition molecules which could ligate with one another during the action of the biomolecular automaton.

In the next part of the algorithm, we will select elements of the family $\mathcal{P}(A_1)$ as sets meant to serve to check the possibilities of generating 40 symbols for the biomolecular automaton using the restriction enzyme BbvI. Thus, let the set C be a chosen element of the family $\mathcal{P}(A_1)$.

In the first part of the stage of deployment and verification, we select a single assignment of 120 words being the elements of the set C, to three sets G_1, G_2 and G_3 (40 words to each set) from among successive possible combinations of assigning the 120 words of the set C to 3 sets consisting of 40 each.

In the second part of the stage of deployment and verification we pre-check whether we are able to form 40 words of length 6 (whether we can create 40 strands in the direction 5′-3′). We examine this by comparing the elements of: first, the sets G_1, G_2 and then G_2, G_3 in the following way:

1. for the sets G_1 and G_2 we check whether the number of occurrences of each word x of length 3, being a suffix in the words of the set G_1, is identical with the

Table 3: Part II: 120 pairs (x,y) of four-element sequences of nucleotides.

No	x	y	No	x	y	No	x	y
61	CAGA	TCTG	81	CGGA	TCCG	101	GCAC	GTGC
62	CAGC	GCTG	82	CGGC	GCCG	102	GCCA	TGGC
63	CAGG	CCTG	83	CGTA	TACG	103	GCCC	GGGC
64	CATA	TATG	84	CGTC	GACG	104	GCGA	TCGC
65	CATC	GATG	85	CTAA	TTAG	105	GCTA	TAGC
66	CCAA	TTGG	86	CTAC	GTAG	106	GGAA	TTCC
67	CCAC	GTGG	87	CTCA	TGAG	107	GGAC	GTCC
68	CCAG	CTGG	88	CTCC	GGAG	108	GGCA	TGCC
69	CCCA	TGGG	89	CTGA	TCAG	109	GGGA	TCCC
70	CCCC	GGGG	90	CTGC	GCAG	110	GGTA	TACC
71	CCCG	CGGG	91	CTTA	TAAG	111	GTAA	TTAC
72	CCGA	TCGG	92	CTTC	GAAG	112	GTCA	TGAC
73	CCGC	GCGG	93	GAAA	TTTC	113	GTGA	TCAC
74	CCTA	TAGG	94	GAAC	GTTC	114	GTTA	TAAC
75	CCTC	GAGG	95	GACA	TGTC	115	TAAA	TTTA
76	CGAA	TTCG	96	GACC	GGTC	116	TACA	TGTA
77	CGAC	GTCG	97	GAGA	TCTC	117	TAGA	TCTA
78	CGAG	CTCG	98	GAGC	GCTC	118	TCAA	TTGA
79	CGCA	TGCG	99	GATA	TATC	119	TCCA	TGGA
80	CGCC	GGCG	100	GCAA	TTGC	120	TGAA	TTCA

(A) 5′-[35]bp GCAGCT -3′
3′-[35]bp CGTCGAAAAA -5′

(B) 5′-[35]bp GCAGCT -3′
3′-[35]bp CGTCGATTTT -5′

(C) 5′-[35]bp GCAGCTTTTTAGCTGC[35]bp -3′
3′-[35]bp CGTCGAAAAATCGACG[35]bp -5′

Figure 11: (A) Transition molecule T_{NS_2} using the sticky end: AAAA. (B) Transition molecule T_{NS_3} using the sticky end: TTTT. (C) Double-stranded fragment of DNA formed as a result of ligation of the two transition molecules presented in (A) and (B).

number of occurrences of the word x as a prefix in the words of the set G_2.

2. for the sets G_2 and G_3 we check whether the number of occurrences of each word x of length 3 being a suffix in the words of the set G_2 is identical with the number of occurrences of the word x as a prefix in the words of the set G_3.

At the stage of generating we introduce the auxiliary set $D = \emptyset$ and examine the possibility of forming 40 words of length 6 (40 strands of symbols in the direction 5′-3′) making use of the elements of the sets G_1, G_2 and G_3, as well as synthesizable concatenations of words of length 3. Each word of length 6 is obtained through a double use of synthesizable concatenations of two words of length 3:

1. we select one word from each of the three sets G_1, G_2 and G_3 in such a way as to make possible concatenation of synthesizable words $x \in G_1$, $y \in G_2$ of length 3 and also to enable concatenation of synthesizable words $y \in G_2$, $z \in G_3$ of length 3.

2. having selected the words $x \in G_1$, $y \in G_2$, $z \in G_3$ which satisfy the above-mentioned condition, we form a word u of length 6 (a strand of symbol in the direction 5′-3′): $u = [[x, y]_3, z]_3$ (symbol assembling, see Fig. 12),

3. the word u obtained upon satisfying the above-presented condition is added to the set D.

In the case where it is impossible to form 40 words (40 words u) from the elements of the sets G_1, G_2 and G_3 in the way given above, we return to checking another possibility of assigning the elements of the set C to the sets G_1, G_2 and G_3. In the case where all the possible assignments of the elements of the set C to the sets G_1, G_2 and G_3, we return to examining the next element of the family $\mathcal{P}(A_1)$. In the case where all the elements of the family

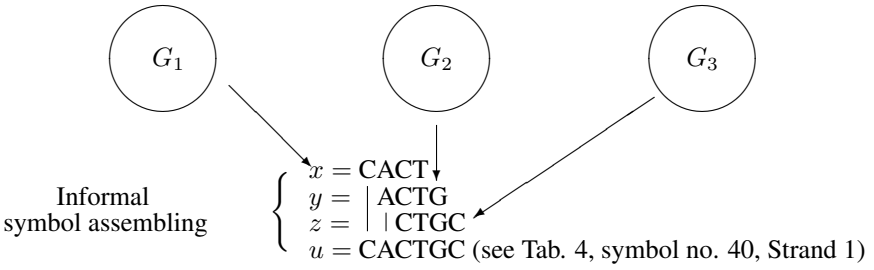


Figure 12: Idea of symbol assembling.

$\mathcal{P}(A_1)$ have been checked and it is impossible to obtain 40 symbols, the algorithm communicates: "Unable to obtain 40 symbols for the biomolecular automaton using the restriction enzyme BbvI".

If the set D has 40 words determined from the elements of the set G_1, G_2 and G_3, we check whether the words of this set (the strands of the symbols in the direction 5′-3′) avoid each of the four, described below, undesired situations, due to the appearance of the sequence recognized by the restriction enzyme BbvI.

The first undesired situation concerns an inclusion of a sequence recognized by the restriction enzyme BbvI inside any symbol. An example to illustrate the above undesired situation is presented in Fig. 13A. Let us note that the second, analogous, undesired situation can occur if a sequence recognized by the restriction enzyme BbvI is included inside any symbol "reversed by 180^o". An example of the latter is shown in Fig. 13B.

(A) 5′-**GCAGC**G-3′ / 3′-CGTCGC-5′ (B) 5′-AGCTGC-3′ / 3′-T**CGACG**-5′

Figure 13: Undesired situations: (A) a sequence recognized by the restriction enzyme BbvI is included in the symbol. (B) a sequence recognized by the restriction enzyme BbvI is contained in the symbol "reversed by 180^o".

The third undesired situation concerns an inclusion of a sequence recognized by the restriction enzyme BbvI in connection of two symbols. An instance illustrating the above-described situation is presented in Fig. 14A. Let us note that the fourth, analogous, undesired situation can occur if a sequence recognized by the restriction enzyme BbvI is included in the connection of two symbols "reversed by 180^o". An example to illustrate the above undesired situation is shown in Fig. 14B.

The appearance of any of the four undesired situations can lead to the occurrence of an undesired action of a biomolecular automaton. In connection with these situations, it is necessary to examine, respectively:

1. whether each word $x \in D$ satisfies the condition: $\sim (e_x \subseteq x)$,
2. whether each word $x \in D$ satisfies the condition: $\sim ((\Xi(e_x))^{-1} \subseteq x)$, where $\sim ((\Xi(e_x))^{-1} \subseteq x)$ means that a sequence recognized by the restriction enzyme BbvI cannot be included in the symbol "reversed by 180^o" and relativized solely to one considered strand in the direction 5′-3′,
3. whether the concatenation $z = xy$ of any words $x \in D$ and $y \in D$ satisfies the condition: $\sim (e_x \subseteq z)$,
4. whether the concatenation $z = xy$ of any words $x \in D$ and $y \in D$ satisfies the condition: $\sim ((\Xi(e_x))^{-1} \subseteq z)$, where $\sim ((\Xi(e_x))^{-1} \subseteq z)$ means that a sequence recognized by the restriction enzyme BbvI cannot be included in the ligation of two symbols "reversed by 180°" and relativized only to one considered strand in the direction 5′-3′.

(A) 5′-TAG**GCAGC**TTAT-3′ / 3′-ATCCGTCGAATA-5′

(B) 5′-TCCGCTGCGGGG-3′ / 3′-AGG**CGACG**CCCC-5′

Figure 14: Undesired situations: (A) a sequence recognized by the restriction enzyme BbvI is included in the ligation of two symbols. (B) a sequence recognized by the restriction enzyme BbvI is included in ligation of two symbols "reversed by 180^o".

In the case one of the four undesired situation is detected, we return to checking another possibility of assigning the elements of the set C to the sets G_1, G_2 and G_3. When all the possible assignments of the elements of the set C to the sets G_1, G_2 and G_3 have been checked, we return to examining another element of the family $\mathcal{P}(A_1)$. In the case all the elements of the family $\mathcal{P}(A_1)$ have been checked and it has been found that it is impossible to obtain 40 symbols that do not include undesired situations, the algorithm returns the message: "Unable to obtain 40 symbols for a biomolecular automaton using the restriction enzyme BbvI". If the set D has 40 words formed from the elements of the sets G_1, G_2, G_3 and there does not occur a single undesired situation, then we move on to the last stage.

At the last stage we determine elements of 40 complementary words for each word of the set D, making use of

Table 4: List of 40 symbols (Strand 1 with its complementary Strand 2) consisting of 6 bp obtained for a biomolecular automaton using restriction enzyme BbvI.

No	Strand 1	Strand 2	No	Strand 1	Strand 2
1	TCGCTA	AGCGAT	21	GCCCGC	CGGGCG
2	CGTTCG	GCAAGC	22	CGCCAG	GCGGTC
3	ATTGAT	TAACTA	23	ATGGGT	TACCCA
4	CGAGTA	GCTCAT	24	GGTAGG	CCATCC
5	CAGGGG	GTCCCC	25	AGGTTA	TCCAAT
6	TAGATA	ATCTAT	26	GCTGTG	CGACAC
7	ATAGTT	TATCAA	27	GTGTAT	CACATA
8	GTTTTG	CAAAAC	28	TATTTA	ATAAAT
9	TTGTTG	AACAAC	29	TTACGA	AATGCT
10	TTGGTG	AACCAC	30	CGACTT	GCTGAA
11	GTGCCG	CACGGC	31	CTTCCG	GAAGGC
12	CTAATG	GATTAC	32	CCGTCT	GGCAGA
13	ATGCGT	TACGCA	33	TCTTAT	AGAATA
14	CGTGAG	GCACTC	34	TATGTC	ATACAG
15	GAGCAA	CTCGTT	35	GTCATC	CAGTAG
16	CAAGCC	GTTCGG	36	ATCGGT	TAGCCA
17	GCCTTT	CGGAAA	37	GGTCCT	CCAGGA
18	TTTCTG	AAAGAC	38	CCTCTC	GGAGAG
19	CTGAAT	GACTTA	39	CTCCAC	GAGGTG
20	AATCCC	TTAGGG	40	CACTGC	GTGACG

the function Ξ of complementarity of words. In this way we acquire pairs of words which mean: a strand of the symbol in the direction $5'$-$3'$ and a strand of the symbol in the direction $3' - 5'$, respectively. On the basis of the presented conception of the algorithm, there were generated 40 words (strands in the direction $5'$-$3'$) denoted as Strand 1 in Tab. 4, as well as 40 words (strands in the direction $3' - 5'$) denoted as Strand 2 in Tab. 4. At the same time, this is giving an answer to the open question asked in 2005: It is possible to generate 40 symbols for a biomolecular automaton using the restriction enzyme BbvI, in which the symbols are coded by means of 6 pairs of nucleotides.

4 Conclusions

The considerations developed in this work aim, on the one hand, to give an answer to the open question posed in the work of Soreni and co-workers (Soreni et al. 2005), relating to the possibility of indicating 40 symbols for a biomolecular automaton which makes use of the restriction enzyme BbvI, in which symbols are coded by means of 6 pairs of nucleotides. On the other hand, they point to the possibility of characterizing the idea of acting of an algorithm which makes it possible to generate 40 symbols for a biomolecular automaton using the restriction enzyme BbvI. Let us note that the open question posed by Soreni and co-workers (Soreni et al. 2005) relating to the possibility of obtaining 40 symbols for a biomolecular automaton using the restriction enzyme BbvI can be generalized in three possible ways: (1) as symbols coded with the use of a different number of pairs of nucleotides, (2) as any other restriction enzyme (used in a biomolecular automaton) and also (3) as the possibility of using more than one restriction enzyme in a biomolecular automaton. Thus, to point to the possibilities of generalization of the question one can start pondering over: (1) the possibility of generating the maximal number of symbols (coded by n pairs of nucleotides) for a biomolecular automaton using one restriction enzyme, (2) the possibility of generating the maximal number of symbols (coded by n pairs of nucleotides) for a biomolecular automaton using more than one restriction enzyme, and also (3) the possibility of an algorithmic approach in each of the two indicated cases. In this way it is possible to raise two general problems which are relativized to the number of restriction enzymes used in a biomolecular automaton and require working out relevant algorithms. Problem 1: generate the maximal number of symbols (coded by n pairs of nucleotides) for a biomolecular automaton using one restriction enzyme. Problem 2: generate the maximal number of symbols (coded by n pairs of nucleotides) for a biomolecular automaton using more than one restriction enzyme. The above-posed problems require considering and defining the conditions which must be imposed on the relations between the restriction enzyme, symbols and other elements which are components of a biomolecular automaton. The output conditions which ought to be considered and taken account of in the above-mentioned relation are the conditions included in the works Krasiński et al. (2013) and Sakowski et al. (2017). Taking into account these conditions will make it possible to determine all the indispensable conditions which serve to elaborate on algorithms enabling to solve the both general problems mentioned above. The solution to the mentioned general problems make it possible to algorithms development for the generating symbols, which are important for laboratory implementation of biomolecular automata.

References

[1] Adleman, L. (1994). Molecular computation of solutions to combinatorial problems. Science, 226, 1021-1024.
https://doi.org/10.1126/science.7973651

[2] Benenson, Y., Paz-Elizur, T., Adar, R., Keinan, E., Livneh, Z., & Shapiro, E. (2001). Programmable and autonomous computing machine made of biomolecules. Nature, 414, 430-434.
https://doi.org/10.1038/35106533

[3] Benenson, Y., Adar, R., Paz-Elizur, T., Livneh, Z., & Shapiro, E. (2003). DNA molecule provides a computing machine with both data and fuel. PNAS, 100, 2191-2196.
https://doi.org/10.1073/pnas.0535624100

[4] Benenson, Y., Gil, B., Ben-Dor, U., Adar, R., Shapiro, E. (2004). An autonomous molecular computer for logical control of gene expression. Nature, 429, 423–429.
https://doi.org/10.1038/nature02551

[5] Bennett, Ch. (1982). The Thermodynamics of computation – a Review. International Journal of Theoretical Physics, 21(12), 905-940.
https://doi.org/10.1007/BF02084158

[6] Bennett, Ch., & Landauer, R. (1985). The fundamental physical limits of computation. Scientific American, 253, 48–56.
https://doi.org/10.1038/scientificamerican0785-48

[7] Chen, P., Jing, L., Jian, Z., Lin, H., Zhizhou, Z. (2007). Differential dependence on DNA ligase of type II restriction enzymes: a practical way toward ligase-free DNA automaton. Biochem. and Bioph. Research Communications, 353, 733-737.
https://doi.org/10.1016/j.bbrc.2006.12.082

[8] Feynman, R. P. (1961). There's plenty of room at the bottom, In D. Gilbert (Ed.) Miniaturization, Reinhol, 282–296.

[9] Gopinath, A., Miyazono, E., Faraon, A., Rothemund, P.W.K. (2016). Engineering and mapping nanocavity emission via precision placement of DNA origami. Nature, 535, 401-405.
https://doi.org/10.1038/nature18287

[10] Krasiński, T., Sakowski, S., Waldmajer, J., Popławski, T. (2013). Arithmetical analysis of biomolecular finite automaton. Fundamenta Informaticae, 128, 463-474.
https://doi.org/10.3233/FI-2013-953

[11] Ran, T., Douek, Y., Milo, L., & Shapiro, E. (2012). A programmable NOR-based device for transcription profile analysis. Scientific reports, 2, 641.
https://doi.org/10.1038/srep00641

[12] Rothemund P. W. K. (1995). DNA and restriction enzyme implementation of Turing machines. Discrete Mathematics and Theoretical Computer Science, 27, 75-120.
https://doi.org/10.1090/dimacs/027/06

[13] Rothemund, P. W., Papadakis, N., & Winfree, E. (2004). Algorithmic self-assembly of DNA Sierpinski triangles. PLoS biology, 2(12), 2041-2053.
https://doi.org/10.1371/journal.pbio.0020424

[14] Rothemund, P.W.K (2006). Folding DNA to Create Nanoscale Shapes and Patterns. Nature, 440, 297-302.
https://doi.org/10.1038/nature04586

[15] Seeman, N. (2001). DNA Nicks and Nodes and Nanotechnology. Nano Letters, 1, 22-26.
https://doi.org/10.1021/nl000182v

[16] Sakowski, S., Krasiński, T., Sarnik, J., Blasiak, J., Waldmajer, J., Poplawski, T. (2017). A detailed experimental study of a DNA computer with two endonucleases. Zeitschrift für Naturforschung C, 72(7-8), 303-313.
https://doi.org/10.1515/znc-2016-0137

[17] Sakowski, S., Krasinski, T., Waldmajer, J., Sarnik, J., Blasiak, J., & Poplawski, T. (2017). Biomolecular computers with multiple restriction enzymes. Genetics and molecular biology, 40(4), 860-870.
https://doi.org/10.1590/1678-4685-gmb-2016-0132

[18] Soreni, M., Yogev, S., Kossoy E., Shoham Y., Keinan E. (2005). Parallel biomolecular computation on surfaces with advanced finite automata. Journal of the American Chemical Society 127, 3935-3943.
https://doi.org/10.1021/ja047168v

[19] Unold, O., Troć, M., Dobosz, T., Trusiewicz, A. (2004). Extended molecular computing model. WSEAS Transactions on Biology and Biomedicine 1, 15-19.

[20] Waldmajer, J., Bonikowski, Z., Sakowski, S. (2019). Theory of tailor automata. Theoretical Computer Science 785, 60-82.
https://doi.org/10.1016/j.tcs.2019.02.002

[21] Whitesides, G. M., Mathias, J. P., & Seto, C. T. (1991). Molecular self-assembly and nanochemistry: a chemical strategy for the synthesis of nanostructures. Science, 254(5036), 1312-1319.
https://doi.org/10.1126/science.1962191

https://doi.org/10.31449/inf.v43i4.2629 Informatica **43** (2019) 495–506 **495**

AMF-IDBSCAN: Incremental Density Based Clustering Algorithm Using Adaptive Median Filtering Technique

Aida Chefrour and Labiba Souici-Meslati
LISCO Laboratory, Computer Science Department, Badji Mokhtar-Annaba University
PO Box 12, Annaba, 23000, Algeria
Computer Science Department, Mohamed Cherif Messaadia, Souk Ahras, Algeria
E-mail: aida_chefrour@yahoo.fr, labiba.souici@univ-annaba.dz

Keywords: incremental learning, DBSCAN, canopy clustering, adaptive median filtering, f-measure

Received: December 28, 2018

Density-based spatial clustering of applications with noise (DBSCAN) is a fundamental algorithm for density-based clustering. It can discover clusters of arbitrary shapes and sizes from a large amount of data, which contains noise and outliers. However, it fails to treat large datasets, outperform when new objects are inserted into the existing database, remove noise points or outliers totally and handle the local density variation that exists within the cluster. So, a good clustering method should allow a significant density modification within the cluster and should learn dynamics and large databases. In this paper, an enhancement of the DBSCAN algorithm is proposed based on incremental clustering called AMF-IDBSCAN which builds incrementally the clusters of different shapes and sizes in large datasets and eliminates the presence of noise and outliers. The proposed AMF-IDBSCAN algorithm uses a canopy clustering algorithm for pre-clustering the data sets to decrease the volume of data, applies an incremental DBSCAN for clustering the data points and Adaptive Median Filtering (AMF) technique for post-clustering to reduce the number of outliers by replacing noises by chosen medians. Experiments with AMF-IDBSCAN are performed on the University of California Irvine (UCI) repository UCI data sets. The results show that our algorithm performs better than DBSCAN, IDBSCAN, and DMDBSCAN.

Povzetek: V članku je predstavljen nov algoritem AMF-IDBSCAN, izboljšana različica DBSCAN, ki uporablja grozdenje krošenj za zmanjšanje obsega podatkov in tehnike AMF za odpravo hrupa.

1 Introduction

Data mining is an interdisciplinary topic that can be defined in many different ways [1]. In the field of database management industry, data analysis is mainly concerned with a number of large data repositories and aims to identify valid, useful, novel and understandable patterns in the existing data.

Clustering is a principal data finding technique in data mining. It separates a data set into subsets or clusters so that data values in the same cluster have some common characteristics or attributes [2]. It aims to divide the data into groups (clusters) of similar objects [3]. The objects in the same cluster are more identical to each other than to those in other clusters. Clustering is widely used in Artificial Intelligence, Pattern recognition, statistics, and other information processing fields.

Many clustering algorithms have been progressed; they may be divided into the following major categories [4]: hierarchical clustering algorithms (BIRCH, CHAMELEON,..), partitioning algorithms (K-means, K-medoids), density-based algorithms (DBSCAN, OPTICS) and grid-based algorithms (STING, CLIQUE).

The input of a cluster analysis system is a set of samples and a measure of similarity (or dissimilarity) between two samples. The output is a set of clusters that form a partition, or a structure of partitions of the data set. Generally, finding clusters is not a simple task and the current clustering algorithms take much time when they are applied to large databases.

In addition, most of the databases are dynamic in nature, data is inserted and deleted from them frequently. The static clustering does not process this kind of databases that's why the concept of incremental clustering was introduced and used.

The difference between the traditional clustering methods (batch mode) and those of incremental clustering is the ability of the latter to process new data included in the data collection without having to perform a full re-clustering. This allows a dynamic following of updates to the database during clustering.

Incremental learning is a research area that received great attention in recent years since it allows effective reuse of data, fast and pragmatic learning based on context, augmentation of knowledge, learning in dynamic and large databases, exploration and smart decision making [5].

In our research, we are interested in evolving incremental clustering to cluster the data objects which the process of updating an existing set of clusters incrementally rather than mining them from scratch on each database update [6]. Evolving clustering algorithms allow incremental changes to be made both structurally

and parametrically through different data-driven mechanisms [7].

In our study, we focus on the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm. The core idea of DBSCAN is that each object within a cluster must have a certain number of other objects in its neighborhood. Compared with other clustering algorithms. DBSCAN has many attractive benefits and was used by many researchers in recent years with its several extensions and applications.

We were particularly interested in the incremental version of DBSCAN since (1) it is capable of discovering clusters of random shape; (2) it requires just two parameters and is most inconsiderate to the ordering of the points in the database; (3) it reduces the search space and facilitates an incremental update in the clusters; (4) it is more adaptive to various datasets and data space without some initial information [8] and (5) the DBSCAN with incremental concept saves a lot of time and effort efficiently, whereas static DBSCAN has already suffered from some drawbacks and these problems are mainly faced in dynamic large databases in the existing system [9];

In this paper, we propose an AMF-IDBSCAN algorithm an enhanced version of the DBSCAN. Our algorithm consists of three main phases. After importing the original database, it preprocesses it to prepare the clustering step and to reduce the volume of the dataset using Canopy clustering. Then, the classical DBSCAN algorithm is applied to the results of the first step to produce another database. Next, the incremental DBSCAN algorithm is applied to the incremental dataset. The adaptive median filtering technique is applied to the results of the previous step to remove noise and outliers. Then, the results are compared and the performance is evaluated.

The rest of the paper is organized as follows. In the next section, we survey in brief the literature of enhanced DBSCAN algorithms. In Section 3, we describe in details our contribution to AMF-IDBSCAN. Section 4 describes the experiment we conducted and the results obtained by our algorithm. It also compares them with top-ranked algorithms. Finally, we draw some conclusions and show ongoing research aspects in Section 5.

2 Related work

Several algorithms for improvements of DBSCAN exist in the literature. In this section, we outline the best known and most recent ones. We noticed that all of these algorithms have shown good results, in the last few years. However, no one of them could be said to be the best but all depend on the content of input parameters and their application domain:

DVBSCAN (A Density-Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases) [10] is an algorithm which handles local density variation within the cluster. The following input parameters are introduced: minimum objects (μ), radius, and threshold values (α, λ), to calculate the growing cluster density mean and the cluster density variance for any core object, which appears to be developed further by considering the density of its Neighborhood with respect to cluster density mean. A comparison between a cluster density variance for a core object and a threshold value is affected if the first is less than the second and is also satisfying the cluster similarity index, and then it will allow the core object for expansion.

ST-DBSCAN (Spatial-Temporal Density-Based Clustering) [11] is constructed to improve the DBSCAN algorithm by introducing the ability to discover clusters with respect to spatial, non-spatial, and temporal values of the objects. ST-DBSCAN works in three stages: (1) It can cluster spatial-temporal data according to spatial, non- spatial, and temporal attributes. (2) To resolve the problem of no detection of the noise input in DBSCAN, ST-DBSCAN assigns density factor to each cluster. (3) To solve the conflicts in border objects, it compares the average value of a cluster with new coming value.

VDBSCAN (Varied Density-Based Spatial Clustering of Applications with Noise) [12] is a new improvement to DBSCAN, it detects cluster with a varied density as well as automatically selects several values of input parameter *Eps* for different densities. It has a two-step procedure. In the first step, the values of *Eps* are calculated for different densities according to a K-dist. plotting. These calculated values are then further used to analyze the clusters with different densities. In the second step, the DBSCAN algorithm is applied with the parameter *Eps* values calculated in the previously discussed step. It ensures that all of the clusters with corresponding densities are clustered.

VMDBSCAN (Vibration Method DBSCAN) [13] is designed for modifying the DBSCAN algorithm. Unlike to existing density-based clustering algorithm, it detects the clusters of different shapes, sizes that differ in local density. VMDBSCAN first extracts the “core” of each cluster after applying DBSCAN. Then it “vibrates" points towards the cluster that has the maximum effect on these points.

DMDBSCAN (Dynamic Method DBSCAN) [14] is a new enhancement of DBSCAN which has pointed out that in clusters, generated by DBSCAN, there is wide density variation. Compared to DBSCAN which uses global *Eps.* It has successfully given the method to compute *Eps* automatically for each of the different density levels in the dataset based on *k-dist.* plot. The major success of this technique includes (1) easy interpretation of generated clusters; (2) no limit on the shape of the generated clusters. DMDBSCAN will use the dynamic method to find a suitable value of Eps for each density level of data set.

L-DBSCAN [15] tries to improve the DBSCAN by a hybrid clustering technique, where *l* stands for leaders. It works as follows: (1) it finds the suitable prototypes from the large dataset; (2) and then it uses the clustering methods on these selected prototypes. The leader clustering method is a fast method and it runs in linear time of the input dataset size. In l-DBSCAN, the first two prototypes are derived with the help of the leader clustering method. Afterwards, DBSCAN is applied to

perform density-based clustering on this prototype respectively.

GRIDBSCAN [16] is another important variation of DBSCAN that addresses the issue that exists in most of the density-based clustering algorithms, which is the lack of accurate clustering in the presence of clusters with different densities. It has a three-level mechanism. In the first level, it provides appropriate grids such that density is similar in each grid. In the next level, it merges the cells having the same densities. At this level, the appropriate value of Eps and *MinPts* are also identified in each grid. In the final step, the DBSCAN algorithm is applied to these identified parameters values to obtain the required final number of clusters.

FDBSCAN (Fast Density-Based clustering algorithm for large Database) [17] is an improved version of DBSCAN clustering. This was developed to overcome: (1) its slow speed (slow in comparison due to neighborhood query for each object); (2) and setting the threshold value of the DBSCAN algorithm. The FDBSCAN starts by ordering the dataset object by certain dimensional coordinates. Then it considers a point having a minimal index and retrieves its neighborhood. If this point is demonstrated as a core object then a new cluster is created to label all objects in its neighborhood. In this way, the next unlabeled point is analyzed outside the core object to expand clusters. When all the points are analyzed for clustering then these objects are further passed through a Kernel function. This will ensure the distribution of object as uniform as possible.

MR-DBSCAN [18] is a parallel version of DBSCAN in a MapReduce manner. It provides a method to divide a large dataset into several partitions based on the data dimensions. In the map phase, localized DBSCANs can be applied to each partition in parallel. During a final reduce phase, the results of each partition are then merged. For the overall cost, a partition-division phase is added into DBSCAN. A Cost Balanced Partition division method is used to generate partitions with equal workloads. This parallel extension meets the requirements of scalable execution for handling large-scale data sets and the MapReduce approach makes it suitable for many popular big data analytics platforms like Hadoop MapReduce and ApacheSpark.

M-DBSCAN (Multi-Level DBSCAN*)* [19] is an algorithm where neighborhood is not defined by a constant radius. Instead, the definition of the neighboring radius is performed based on the data distribution around the core using standard deviation and mean values. To obtain the clustering results, M-DBSCAN is applied on a set of core-mini clusters where each core-mini cluster defines a virtual point which lies in the center of that cluster. In M-DBSCAN, the value of DBSCAN is replaced by local density cluster which the clusters are extended by adding core-mini clusters that have similar mean values with a little difference determined by the standard deviation of the core.

FI-DBSCAN [20] is a Frequent Itemset Ultrametric Trees with Density Based Spatial Clustering of Applications with Noise (DBSCAN) on MapReduce framework is used in the proposed system to solve the evolution and efficiency problem in an existing frequent itemset. It incorporates the Density Based Frequent Itemset Ultrametric Tree by adding additional hash tables rather than using conventional FP trees, there are by achieving compressed storage and avoiding the necessity to build conditional pattern bases. FI-DBSCAN integrates three MapReduce jobs to accomplish parallel mining of frequent itemsets. The first MapReduce job is responsible for mining all frequent one- itemsets. The second MapReduce job applies the second round of analyzing the database to eliminate infrequent items from each transaction record. At the end of the third MapReduce job, all frequent K-itemsets are created.

AnyDBC (An Efficient Anytime Density-based Clustering Algorithm for Very Large Complex Datasets) [21] is an anytime algorithm which requires very small initial runtime for acquiring similar results as DBSCAN. Thus, it not only allows user interaction but also can be used to obtain good approximations under arbitrary time constraints.

IDBSCAN [22] proposes an enhanced version of the incremental DBSCAN algorithm for incrementally building and updating arbitrarily shaped clusters in extensive datasets. The proposed algorithm ameliorates the incremental clustering process by limiting the search space to partitions instead of the whole dataset, and this gives significant improvements in performance compared to relevant incremental clustering algorithms. To enhance this algorithm further, [23] proposes an incremental DBSCAN which is fused with a suitable noise removal and outlier detection technique inspired by the box plot method. It utilizes a between network measure to dense regions to frame the last number of clusters.

3 The proposed AMF-IDBSCAN clustering algorithm

To overcome the limitations of the high complexity and the non scalability of the traditional clustering algorithms, we have developed in this work AMF-DBSCAN: An enhanced incremental DBSCAN using a canopy clustering algorithm and an adaptive median filtering technique.

The proposed AMF-IDBSCAN consists of four phases as shown in Figure 1. The first phase is pre-clustering employing Canopy clustering. The second phase is the clustering of data objects in which Incremental DBSCAN is used. The third phase is post-clustering applying Adaptive Median Filtering method that aims to reduce the number of outliers by replacing them with chosen medians. The last phase is used to evaluate the performance of clustering algorithms using different evaluation metrics:

In the next subsections, we describe in details the main steps of our algorithm.

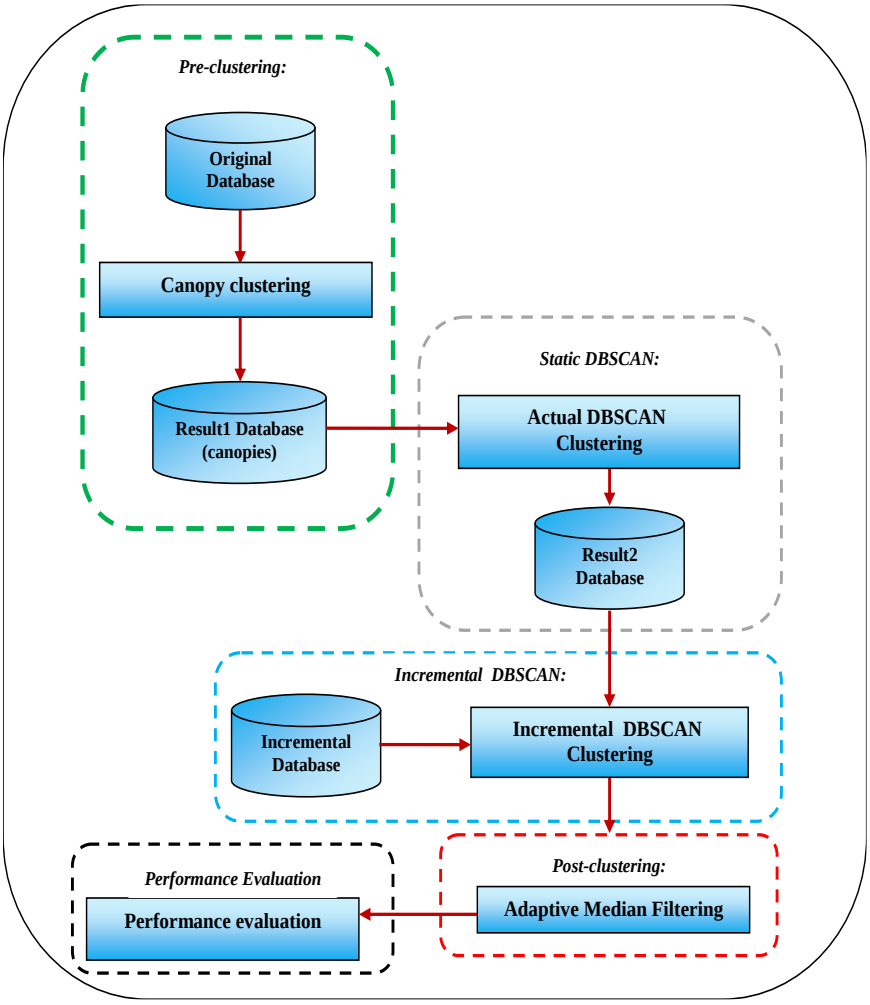


Figure 1: The methodology of the proposed incremental AMF-IDBSCAN clustering algorithm.

3.1 Pre-clustering

This step is aims to prepare the clustering. We used the canopy clustering algorithm which is an unsupervised pre-clustering algorithm introduced by [24]. We have chosen a canopy clustering method for pre-processing the data because (1) it is efficient when the problem is large (2) it can greatly reduce the number of distance computations required for clustering by first cheaply partitioning the data into overlapping subsets, and then only measuring distances among pairs of data points that belong to a common subset and (3) it tries to speed up the clustering of large data sets that are a high dimension by dividing the clustering process into two subprocesses, where using another algorithm directly may be impractical due to the size of the data set (see Figure 2).

First, the data set is divided into overlapping subsets called canopies. This is done by choosing a distance metric and two thresholds, T1 and T2, where T1 > T2.

All data points are then added to a list and one of the points in the list is picked at random. The remaining points in the list are iterated over and the distance to the initial point is calculated. If the distance is within T1, the point is added to the canopy. Further, if the distance is within T2, the point is removed from the list. The algorithm is iterated until the list is empty.

The output of the Canopy clustering is the input of static DBSCAN;

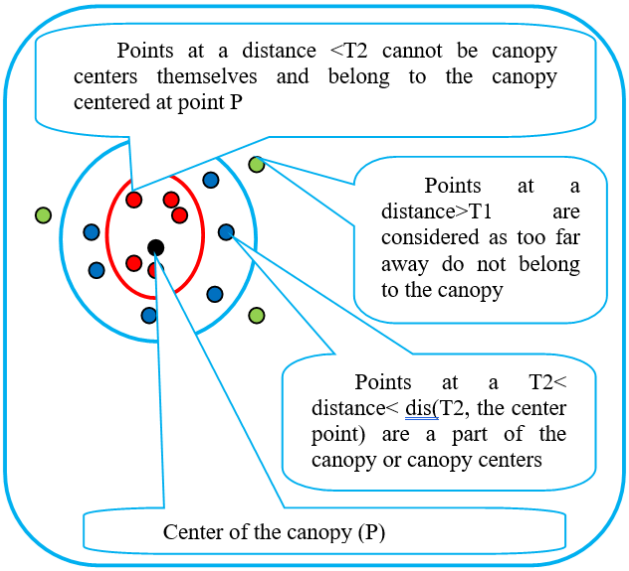


Figure 2: Canopy clustering description [25].

3.2 Classical static DBSCAN clustering algorithm

When used with canopy clustering, the DBSCAN algorithm can reduce the computations in the radius (Eps) calculation step that delimits the neighborhood area of a point hence improving the efficiency of the algorithm. The implementations of the DBSCAN algorithm with Canopy Clustering involves the following steps (see Figure 3):

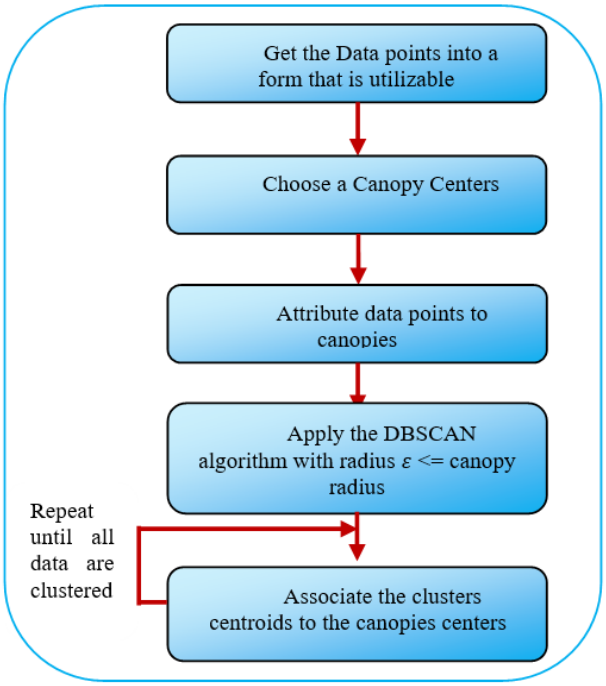


Figure 3: DBSCAN algorithm with Canopy Clustering.

1. Prepare the data points: the input data needs to be transformed into a format suitable and utilizable for distance and similarity measures.
2. Choose Canopy Centers
3. Attribute data points to canopy centers: the canopy assignment step would simply assign data points to generated canopy centers.
4. Associate the cluster's centroids to the canopies centers. The data points are now in clustered sets.
5. Repeat the iteration until all data are clustered.
6. Apply the DBSCAN algorithm with radius ε<= canopy radius and iterate until clustering. The computation to calculate the minimum number of points (Minpts) is greatly reduced as we only calculate the distance between a clusters centroids and data point if they share the same canopy.
7. DBSCAN is a widely used technique for clustering in spatial databases. DBSCAN needs less knowledge of input parameters. The major advantage of DBSCAN is to identify arbitrary shape objects and removal of noise during the clustering process. Besides its familiarity, it has problems with handling large databases and in the worst case, its complexity reaches to $O(n^2)$[26]. Additionally, DBSCAN cannot produce a correct result on varied densities. That's why we used canopy clustering in our case to reduce its complexity:

In the AMF-IDBSCAN algorithm, we partition the data (n is the number of data) into canopies C by canopy clustering, each containing about (n / C) points. Then the complexity will decrease to (n^2 /C) for the AMF-IDBSCAN algorithm.

In the sub-section, we describe the static DBSCAN:

The static DBSCAN algorithm was first introduced by [27]. It uses a density-based notion of clustering of arbitrary shapes, which is designed to discover clusters of arbitrary shape and also has the ability to handle noise. It relies on the density-based notion of clusters. Clusters are identified by looking at the density of points.

Regions with a high density of points depict the existence of clusters, whereas regions with a low density of points indicate clusters of noise or clusters of outliers.

The key idea of the DBSCAN algorithm [28] is that, for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points, that is, the density in the neighborhood has to exceed some predefined threshold. This algorithm needs two input parameters:

Eps, the radius that delimits the neighborhood area of a point (Eps-neighborhood);

MinPts, the minimum number of points that must exist in the Eps-neighborhood.

The clustering process is based on the classification of the points in the dataset as core points, border points, and noise points, and on the use of density relations between points (directly density-reachable, density-reachable, density-connected) to form the clusters (see Figure 4).

Core point: lies in the interior of density based clusters and should lie within Eps (radius or threshold value), MinPts (minimum number of points) which are user-specified parameters.

Border point: lies within the neighborhood of core point and many core points may share the same border point.

Noise point: is a point which is neither a core point nor a border point.

Directly Density-Reachable: a point P is directly density-reachable from a point Q with respect to (w.r.t) Eps, MinPts if P belongs to NEps(Q) |NEps (Q)| >= MinPts

Density-Reachable: a point P is density-reachable from a point Q w.r.t Eps, MinPts if there is a chain of points P_1, …, P_n, P_1 = Q, P_n = P such that P_{i+1} is directly density-reachable from P_i

Density-Connected: a point P is density-connected to a point Q w.r.t Eps, MinPts if there is a point O such that both, P and Q are dense-reachable from O w.r.t Eps and MinPts.

The steps of the DBSCAN algorithm are as follows [27]:

- Arbitrary select a point P
- Retrieve all points density-reachable from P w.r.t Eps and MinPts.
- If P is a core point, a cluster is formed.
- If P is a border point, no points are dense-reachable from P and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

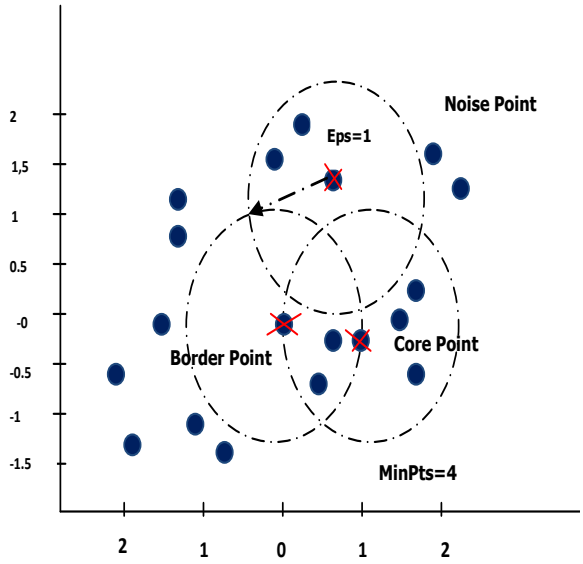


Figure 4: DBSCAN working.

3.3 Incremental DBSCAN clustering

The static DBSCAN approach is not suitable for a large multidimensional database which is frequently updated. In that case, the incremental clustering approach is much finer. In our study, we use incremental DBSCAN to enhance the clustering process by incrementally partitioning the dataset to reduce the search space of the neighborhood to one partition rather than the whole data set. Also, it has embedded flexibility regarding the level of granularity and is robust to noisy data.

We have chosen as a foundation of our incremental DBSCAN clustering algorithm, the algorithm of [29] which works in two steps:

Step 1. Compute the means between every core object of the clusters and the new data. Insert the new data into a specific cluster based on the minimum mean distance. Sign the data as noise or border if it cannot be inserted into any clusters.

Step 2. Form new core points or clusters when noise points or border points fulfill the Minpts (the minimum number of points) and radius criteria.

Sometimes DBSCAN may be applied on a dynamic database which is frequently updated by the insertion or deletion of data. After insertions and deletions to the database, the clustering located by DBSCAN has to be updated. Incremental clustering could enhance the chances of finding the global optimum. In this approach, first, it will form clusters based on the initial objects and a given radius (eps) and Minpts. Thus the algorithm finally gets some clusters fulfilling the conditions and some outliers. When new data is inserted into the existing database, we have to update the existing clusters using DBSCAN. At first, the algorithm computes the means between every core object of clusters and the new coming data and insert it into a particular cluster based on the minimum mean distance. The new data which is not inserted into any clusters, is treated as noise or outlier. Sometimes outliers which fulfill the Minpts & Eps criteria, combined can form clusters using DBSCAN.

We have used the Euclidean distance because it is currently the most frequently used metric space for the established clustering algorithms [30].

The steps of incremental DBSCAN clustering algorithm are as follows:

Pseudo-code of incremental DBSCAN
Input
D: a dataset containing n objects $\{X_1, X_2, X_3 \ldots, X_n\}$; n: number of data items; Minpts: Minimum number of data objects ; eps: radius of the cluster.
Output
K: a Set of clusters.
Procedure

Let, Ci (where i=1, 2, 3 …) is the new data item.
1. Run the actual DBSCAN algorithm and clustered the new data item C_i properly based on the radius(eps) and the Minpts criteria. Repeat till all data items are clustered.
Incremental DBSCAN Pseudo-code:
Start:
2. a> Let, K represents the already existing clusters.
b>When new data is coming into the database, the new data will be directly clustered by calculating the minimum mean(M) between that data and the core objects of existing clusters.
for i = 1 to n do
find some mean M in some cluster K_p in
K such that dis (C_i, M) is the smallest;
If (dis (C_i, M) is minimum) && (C_i<=eps) && (size(K_p)>=Minpts) then
$K_p = K_p$ U C_i ;
Else
If dis (C_i != min) || (C_i>eps) ||(size(K_p)<Minpts) then
C_i Outlier(O_i) .
Else
If Count(O_i) Minpts then O_i Form new cluster(M_i).
C > Repeat step b till all the data samples are clustered.
End.

3.4 Post-clustering

We illustrate the clusters and the outliers points by a rectangular window W on a hyperplane of n dimensions equivalent to the data dimensions. We apply the Adaptive Median Filtering (AMF) to reduce the noise

data. We have taken the key idea of this method and we have applied it to our proposition. This is an important advantage of our approach.

We have selected Adaptive Median Filtering (AMF) among various filtering techniques because it removes noise while preserving shape details [31]. AMF technique is used to replace the outliers generated by incremental DBSCAN by a cluster contains an object.

The adaptive median filtering [32] has been widely applied as an advanced method compared with standard median filtering. The adaptive filter works on a rectangular region W (illustration of the set of clusters and outliers generated by the previous stage on a hyperplane). It changes the size of W during the filtering operation depending on certain criteria as listed below. The output of the filter is a single value which replaces the current noise data value at (x, y,....), the point on which W is centered at the time.

Let $I_{x,y,....}$ be the selected noise data according to the dimensions, I_{min} be the minimum noise value and I_{max} be the maximum noise value in the window, W be the current window size applied, W_{max} be the maximum window size that can be achieved and I_{med} be the median of the window designated. The algorithm of this filtering technique completes in two levels as described in [33]:

Level A:

a) If $I_{min} < I_{med} < I_{max}$ then the median value is not an impulse, so the algorithm goes to Level B to check if the current noise is an impulse.

b) Else the size of the window is increased and Level A is repeated until the median value is not a stimulus so the algorithm goes to Level B; or the maximum window size is reached, in which case the median value is assigned as the filtered selected noise value.

Level B:

a) If $I_{min} < I_{x,y,....} < I_{max}$, then the current noise value is not a stimulus, so the filtered selected noise is unchanged

b) Else the selected noise data is either equal to I_{max} or I_{min}, then the filtered selected noise data is assigned the median value from Level A.

These types of median filters are widely used in filtering data that has been denoised with noise density greater than 20%.

This technique has three main purposes:

- To remove noise;
- To smoothen any non-stimulus noise;
- To reduce excessive shapes of clusters

3.5 Performance evaluation

To evaluate the performance of our approach, the canopies are applied to the original dataset and store the result into another database, and then the actual DBSCAN algorithm is applied to the results to this database. The incremental DBSCAN algorithm is applied to the incremental dataset. The results of these two algorithms are compared and their performances are evaluated.

The proposed algorithm AMF-IDBSCAN is shown as pseudo-code in Algorithm 2:

Pseudo-code of AMF-IDBSCAN

Input

D: a dataset containing n objects $\{X_1, X_2, X_3 \ldots, X_n\}$;
n: number of data items;
Minpts: Minimum number of data objects ;
eps: radius of the cluster.
CN: canopies centers

Output

K: a Set of clusters.
A single value: $I_{x,y,....}$ or I_{med}

Procedure

1. Run Canopy clustering :

1.1. Put all data into a List, and initialize two distance radius about the loose threshold T1 and the tight threshold T2 (T1> T2).

1.2. Randomly select a point as the first initial center of the Canopy cluster, and delete this object from the List.

1.3. Get a point from the List, and calculate the distance d to each Canopy clusters.

If d < T2, the point belongs to this cluster; if T2≤d≤T1, this point will be marked with a weak label;

If the distance d to all Canopy center is greater than T1, then the point will be classed as a new Canopy cluster center. Finally, this point should be eliminated from the List;

1.4. Run the step1.3 repeatedly until the list is empty, and recalculate the canopy centers CN.

2. Run the actual DBSCAN algorithm and clustered the new data item C_i properly based on the radius(eps) and the Minpts criteria. Repeat till all data items are clustered:

2.1. Choosing Canopy Centers CN

2.2. Attribute data points D to canopy centers CN;

2.3. Apply the DBSCAN algorithm with radius ε <= canopy radius with dist(CN, Ci)<Minpts

2.4. Repeat the iteration until all data are clustered.

3. Run the incremental DBSCAN:

3.1. a> Let, K represents the already existing clusters.

3.2. When new data is coming into the database, the new data will be directly clustered by calculating the minimum mean(M) between that data and the core objects of existing clusters.

For i = 1 to n do
find some mean M in some cluster K_p in
K such that dis (C_i, M) is the smallest;
If (dis (C_i, M) is minimum) && (C_i<=eps) && (size(K_p)>=Minpts) then
$K_p = K_p \cup C_i$;
Else
If dis (C_i != min) || (C_i>eps) ||(size(K_p)<Minpts) then
C_i Outlier(O_i).

3.3. Elimination of noise objects Oi:

The new dataset contains Oi and the clusters closest to it.

3.4. Adaptive Median Filtering Technique:

For i=1 to m do {where m is the number of outliers}
Illustrate a new rectangle on a hyperplane;

Let:
$I_{x,y,...}$ be the selected noise data (Oi) at the coordinates (x,y,...); % corresponding the data dimensions;
I_{min} be the minimum noise value;
I_{max} be the maximum noise value in the window;
W be the current window size applied; % It contains K clusters and Oi
W_{max} be the maximum window size that can be reached;
I_{med} be the median of the window assigned
Algorithm
Level A: A1= I_{med} - I_{min}
A2= I_{med} - I_{max}
If A1 > 0 AND A2 < 0, Go to level B
Else increase the window size
If window size W<= I_{max} repeat level A
Else output $I_{x,y..}$.

Level B: B1 = $I_{x,y,..} – I_{min}$
B2 = $I_{x,y,..} – I_{min}$
If B1 > 0 And B2 < 0 output $I_{x,y,..}$
Else output I_{med}.

3.5. Repeat step b till all the data samples are clustered.
4. Evaluate performance.

4 Experiments and results

This section presents a detailed experimental analysis carried out to prove our proposed clustering technique AMF-IDBSCAN is better than other state of art methods used for high dimensional clustering. We have taken five high dimensional data sets (Adult, Wine, Glass identification, Ionosphere, and Fisher's Iris) from UC Irvine repository (refer Table 1) to test the performance in terms of F-measure, number of clusters, error rate, number of uncluttered instances and time is taken to build the model. F-measure is defined in equation (1), it is the harmonic average of precision and recall. It is a one only summary statistic that does not credit an algorithm for correctly placing the very large number of pairs into different clusters [34]. F-Measure is commonly used in evaluating the efficiency and the reliability of clustering and classification algorithms.

Dataset	No. of instances	No. of attributes	Attribute type	Data types
Ionosphere	351	34	Integer, real	Multivariate
Wine	178	13	Integer, real	Multivariate
Glass Identification	214	10	Real	Multivariate
Adult	48842	15	Categorical, Integer, Real	Multivariate
Fisher's Iris	150	4	Real	Multivariate

Table 1: Description of UCI databases.

Our proposed noise removal and outlier labeling method are compared with static DBSCAN, an incremental density based clustering algorithm (IDBSCAN) [22], DMDBSCAN [14] presented below is the brief related work, about evaluation metrics used for evaluating clustering results:

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (1)$$

$$Where:$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Where TP: True positive, FP: False positive, FN: False negative

DBSCAN:

We apply the DBSCAN algorithm on wine dataset with Eps = 0.9 and MinPts = 6, F-measure= 0.175 and obtain an average error index of 26.18%, number of clusters = 3. While applying DBSCAN on Iris data set, we get an average error index of 35.33% with the same Eps and Minpts, F-measure= 0.264, number of clusters = 3. Another real data set is Glass dataset and when we apply DBSCAN on it, we get an average error index of 68.22 %, F-measure= 0.423 with a number of = 6. We get for Adult dataset an average error index of 32%, F-measure= 0.475 with number of clusters=1216. For ionosphere, an average error index= 31.62%, F-measure= 0.854 and number of clusters=2 (see Table 2)

Dataset Name	N° of clusters	F-measure	Time taken to build a model	Error rate (%)	N° of Unclustered instances
Fisher's iris	3	0.264	0.02	35.33	0
Wine	3	0.175	0.06	26.18	1
Glass identification	6	0.423	0.04	68.22	4
Adult	1216	0.475	1638,35	32	15813
Ionosphere	2	0.854	0.23	31.62	111

Table 2: Results of applying DBSCAN.

IDBSCAN:

We apply IDBSCAN algorithm on wine dataset with Eps = 0.9 and MinPts = 6, F-measure= 0.274 and obtain an average error index of 23.45%, number of clusters = 4. While applying IDBSCAN on Iris data set, we get an average error index of 28.54% with the same Eps and Minpts, F-measure= 0.354, number of clusters = 3. Another real data set is Glass dataset and when we apply IDBSCAN on it, we get an average error index of 49.52%, F-measure= 0.323 with a number of clusters = 8. We get for Adult dataset an average error index of

27.96%, F-measure= 0.475 with number of clusters=1265. For ionosphere, an average error index= 29.15%, F-measure= 0.639 and number of clusters=3 (see Table3).

Dataset Name	N° of clusters	F-measure	Time taken to build a model	Error rate (%)
Fisher's iris	3	0.354	0.03	28.54
Wine	4	0.274	0.05	23.45
Glass identification	8	0.323	0.09	49.52
Adult	1265	0.475	5476.9	27.96
Ionosphere	3	0.639	0.84	29.15

Table 3: Results of applying IDBSCAN.

AMF-IDBSCAN:

In our experiments, we have used for canopy clustering implementation, a Weka tool (Waikato Environment for Knowledge Analysis) [35] which is an open-source Java application produced by the University of Waikato in New Zealand. It functions like Preprocessing Filters, Attribute selection, Classification/Regression, Clustering, Association discovery, Visualization. The set of training instances has to be encoded in an input file with.ARFF (Assign Relation File Format) extension to be used by the Weka tool in order to generate the canopies that will be used as inputs in our algorithm.

We apply our proposed AMF-IDBSCAN algorithm on wine data set with Eps = 0.9 and MinPts = 6, number of canopies= 4, F-measure= 0.354 and obtain an average error index of 18.25%, number of clusters = 4. While applying AMF-IDBSCAN on Iris data set, we get an average error index of 25.63% with the same Eps and Minpts, number of canopies= 3, F-measure= 0.758, number of clusters = 4. Another real data set is Glass data set and when we apply our proposed algorithm on it, we get an average error index of 35.96% , number of canopies= 8, F-measure= 0.695 with number of = 6. We get for Adult dataset an average error index of 29.46%, number of canopies= 100, F-measure= 0.495 with number of clusters=1285. For ionosphere, an average error index= 27.64%, number of canopies= 11, F-measure= 0.821 and number of clusters=5 (see Table 4).

DMDBSCAN:

We apply the DMDBSCAN algorithm on the wine data set, and applying k-dist for 3-nearest points, we have 3 values of Eps which are 4.3, 4.9 and 5.1. F-measure= 0.125, the average error index is 23.15% and number of clusters = 3. While applying DMDBSCAN on Iris data set, and applying k-dist for 3-nearest points, we have 2 values of Eps which are 0.39 and 0.45. The average error index is 38.46%, F-measure =0.295 and a number of clusters = 3.

Another real data set is Glass dataset and when we apply DMDBSCAN on it and applying k-dist for 3-nearest points, we have 3 values of Eps which are 0.89, 9.3 and 9.4. F-measure= 0.623, the average error index is 58.39% and the number of clusters = 6. We get for Adult dataset an F-measure= 0.474, the average error index of 34.66%, with a number of clusters=1301. For ionosphere, F-measure= 0.754, an average error index= 30.04%, and number of clusters=6 (see Table 5).

Dataset Name	T1	T2	N° of canopies	N° of clusters	F-measure	Error rate (%)	Time taken to build a model
Fisher's iris	1.092	0.874	3	4	0.798	25.63	0.01
Wine	1,561	1,249	4	4	0.354	18.25	0.02
Glass identification	1,237	0,989	8	6	0.695	35.96	0.04
Adult	2,020	1,616	100	1285	0.495	29.46	0.07
Ionosphere	2,700	2,160	11	5	0.821	27.64	0.04

Table 4: Results of applying AMF-IDBSCAN.

Dataset Name	N° of clusters	F-measure	Error rate (%)	Time is taken to build a model
Fisher's iris	3	0.293	38.46	0.08
Wine	3	0.125	23.15	0.13
Glass identification	6	0.623	58.39	0.24
Adult	1301	0.474	34.66	0.64
Ionosphere	6	0.754	30.04	0.09

Table 5: Results of applying DMDBSCAN.

Dataset Name	DBSCAN			AMF-IDBSCAN			DMDBSCAN			IDBSCAN		
	N° of clusters	F-measure	Error rate (%)	N° of clusters	F-measure	Error rate (%)	N° of clusters	F-measure	Error rate (%)	N° of clusters	F-measure	Error rate (%)
Fisher's iris	3	0.264	35.33	4	0.798	25.63	3	0.293	38.46	3	0.354	28.54
Wine	3	0.175	26.18	4	0.354	18.25	3	0.125	23.15	4	0.274	23.45
Glass identification	6	0.423	68.22	6	0.695	35.96	6	0.623	58.39	8	0.323	49.52
Adult	1216	0.475	32	1285	0.495	29.46	1301	0.474	34.66	1265	0.475	27.96
Ionosphere	2	0.854	31.62	5	0.821	27.64	6	0.754	30.04	3	0.639	29.15

Table 6: Comparison against the results of DBSCAN, IDBSCAN, DMDBSCAN and our proposed algorithm AMF-IDBSCAN.

Table 6 compares the results obtained by our proposed algorithm against those of three other algorithms, namely: DBSCAN, IDBSCAN, and DMDBSCAN:

- From our experiments, and as Tables 2, 3, 4 and 5 show: by using the DBSCAN algorithm for multi-densities data sets, we get low-quality results with long times. DBSCAN algorithm is a time-consuming algorithm when dealing with large datasets. This is due to Eps and Minpts parameters values which are very important for DBSCAN algorithm, but their calculations are time-consuming. In other sense, clustering algorithms are in need to discover a better version of the DBSCAN algorithm to deal with these special multi-densities datasets.
- DMDBSCAN gives better efficiency results than DBSCAN clustering algorithm but takes more time compared with the other algorithms. This is due that the algorithm needs to call the DBSCAN algorithm for each value of Eps.
- The IDBSCAN algorithm is more efficient in terms of error rate and f-measure than DBSCAN algorithm. Also, it takes more time compared with DBSCAN, DMDBSCAN, and AMF-IDBSCAN. This is due to the fact that this algorithm needs to call the DBSCAN algorithm to make the initial clustering.
- AMF-IDBSCAN gives the best efficiency results compared to the other studied algorithms. Table 6 presents the F-Measure values recorded for all the data sets and all the algorithms. A high value of F-Measure proves the better quality of the clustering process. A significant improvement is found on AMF-IDBSCAN and on all datasets except the Ionosphere dataset. The maximum increase is observed in both Iris and Glass data sets. The improvement in F-Measure shows that our proposed method is more efficient in terms of noise removal and outlier labeling. Apart from F-Measure, our proposed method allows to achieve good clustering results in a reasonable time.

  It can be easily observed from Figure 5 that our proposed clustering method for noise removal is well suited for high dimensional data sets and it exceeds the other existing methods.

5 Conclusion and perspectives

In this paper, we proposed AMF-IDBSCAN an enhanced version of the DBSCAN algorithm, including the notions of density, canopies and noise removal. This work presents a comparative study of the performance of this proposed approach which is fused with an adaptive median filtering median for noise removal and outlier detection technique and a canopy clustering method to reduce the volume of large datasets.

We compared this algorithm with the original DBSCAN algorithm, IDBSCAN, DMDBSCAN, and our experimental results show that the proposed approach gives better results in terms of error rate and f-measure with the increment of data in the original database.

In our future works, we will extend our investigations to other incremental clustering algorithms like COBWEB, incremental OPTICS and incremental supervised algorithms like incremental SVM, learn++, etc.

One of the remaining interesting challenges is how to select the algorithm parameters like k-dist, eps, Minpts, and number of canopies automatically.

6 Acknowledgment

The authors are grateful to the anonymous referees for their very constructive remarks and suggestions.

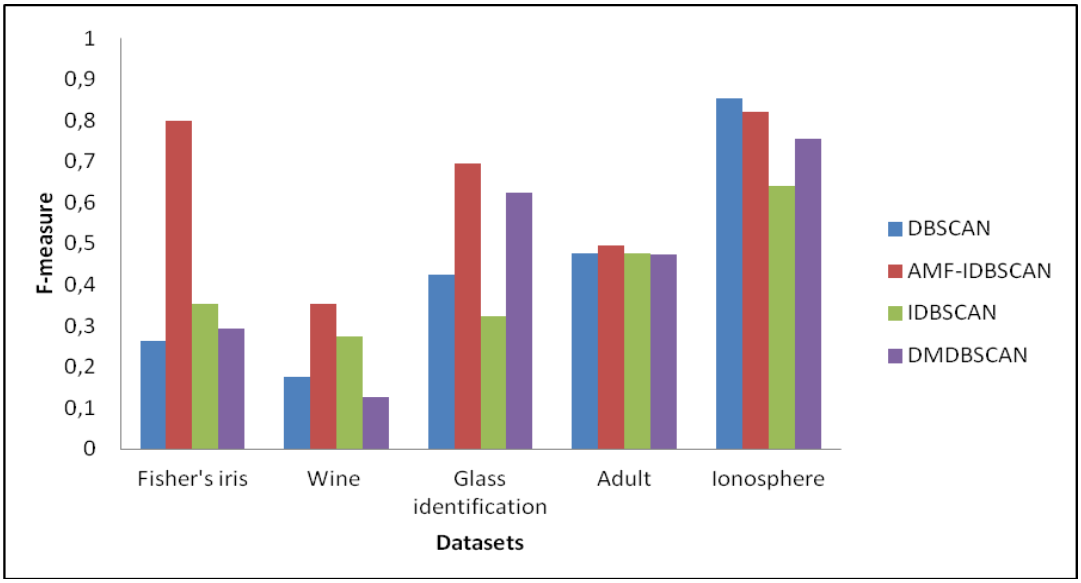


Figure 5: F-Measure (FM) Comparison across all Datasets.

7 References

[1] Patil Y.S., Vaidya M.B. (2012). A technical survey on cluster analysis in data mining. *Journal of Emerging Technology and Advance Engineering*, 2(9):503-513.

[2] ur Rehman S., Khan M.N.A. (2010). An incremental density-based clustering technique for large datasets. *In Computational Intelligence in Security for Information Systems., Springer, Berlin, Heidelberg*,pp. 3-11. http:// doi.org/10.1007/978-3-642-16626-6_1

[3] Abudalfa S., Mikki M. (2013). K-means algorithm with a novel distance measure. *Turkish Journal of Electrical Engineering & Computer Sciences*, 21(6): 1665-1684. https://doi.org/10.3906/elk-1010-869.

[4] Kumar K. M., Reddy A. R. M. (2016). A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method. *Pattern Recognition*, 58:39-48. https://doi.org /10.1016/j.patcog.2016.03.008.

[5] Kulkarni P. A., Mulay P. (2013). Evolve systems using incremental clustering approach. *Evolving Systems*, 4(2): 71-85. https://doi.org /10.1007/s12530-012-9068-z.

[6] Goyal N., Goyal P., Venkatramaiah K., Deepak P. C., Sanoop P. S. (2011, August). An efficient density based incremental clustering algorithm in data warehousing environment. *In 2009 International Conference on Computer Engineering and Applications, IPCSIT* , 2: 482-486. https://doi.org/10.1016/j.aej.2015.08.009.

[7] Tseng F., Filev D., Chinnam R. B. (2017). A mutual information based online evolving clustering approach and its applications. *Evolving Systems*, 8(3): 179-191. https://doi.org/10.1007/s12530-017-9191-y.

[8] Liu X., Yang Q., He L. (2017). A novel DBSCAN with entropy and probability for mixed data. *Cluster Computing*, 20(2):1313-1323. https://doi.org/10.1007/s10586-017-0818-3.

[9] Suthar N., Indr P., Vinit P. (2013). A Technical Survey on DBSCAN Clustering Algorithm. *Int. J. Sci. Eng. Res*, 4:1775-1781.

[10] Ram A., Jalal S., Jalal A. S., Kumar M. (2010). DVBSCAN: A density based algorithm for discovering density varied clusters in large spatial databases. *International Journal of Computer Applications*, pp. 0975-8887. https://doi.org/10.5120/739-1038.

[11] Birant D., Kut A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1): 208-221. https://doi.org/10.1016/j.datak.2006.01.013.

[12] Liu P., Zhou D., Wu N. (2007, June). Varied density based spatial clustering of application with noise. *In International Conference on Service Systems and Service Management*, p. 21. https://doi.org/10.1109/ICSSSM.2007.4280175.

[13] Elbatta M. N. (2012). An improvement for DBSCAN algorithm for best results in varied densities. https://doi.org/10.1109/MITE.2013.6756302.

[14] Elbatta M. T., Ashour W. M. (2013). A dynamic method for discovering density varied clusters, 6(1).

[15] Viswanath P., Pinkesh R. (2006, August). l-dbscan: A fast hybrid density based clustering method. *In 18th International Conference on Pattern Recognition (ICPR'06)*, IEEE, 1: 912-915. https://doi.org/10.1109/ICPR.2006.741.

[16] Uncu O., Gruver W. A., Kotak D. B., Sabaz D., Alibhai Z., Ng C. (2006, October). GRIDBSCAN: GRId density-based spatial clustering of applications with noise. *In 2006 IEEE International Conference on Systems, Man and Cybernetics* , IEEE, 4: 2976-2981.

https://doi.org/10.1109/ICSMC.2006.384571.

[17] Liu B. (2006, August). A fast density-based clustering algorithm for large databases. *In 2006 International Conference on Machine Learning and Cybernetics* IEEE,pp. 996-1000. https://doi.org /10.1109/ICMLC.2006.258531.

[18] He Y., Tan H., Luo W., Mao H., Ma D., Feng S., Fan J. (2011, December). Mr-dbscan: an efficient parallel density-based clustering algorithm using mapreduce. *In 2011 IEEE 17th International Conference on Parallel and Distributed Systems, IEEE,* pp. 473-480. https://doi.org /10.1109/ICPADS.2011.83.

[19] [19] Wang S., Liu Y., Shen B. (2016, July). MDBSCAN: Multi-level density based spatial clustering of applications with noise. *In Proceedings of the 11th International Knowledge Management in Organizations Conference on The changing face of Knowledge Management Impacting Society, ACM,* p. 21. https://doi.org /10.1145/2925995.2926040.

[20] Swathi Kiruthika V., Thiagarasu Dr. V. (2017). FI-DBSCAN: Frequent Itemset Ultrametric Trees with Density Based Spatial Clustering Of Applications with Noise Using Mapreduce in Big Data. *International Journal of Innovative Research in Computer and Communication Engineering*, 5: 56-64. https://doi.org/ 10.15680/IJIRCCE.2017. 0501007

[21] Mai S. T., Assent I., Storgaard M. (2016, August). AnyDBC: an efficient anytime density-based clustering algorithm for very large complex datasets. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM,* pp. 1025-1034. https://doi.org/ 10.1145/2939672.2939750.

[22] Bakr A. M., Ghanem N. M., Ismail M. A. (2015). Efficient incremental density-based algorithm for clustering large datasets. *Alexandria Engineering Journal*, 54(4):1147-1154. https://doi.org/ 10.1016/j.aej.2015.08.009.

[23] Yada P., Sharma P. (2016). An Efficient Incremental Density based Clustering Algorithm Fused with Noise Removal and Outlier Labelling Technique. *Indian Journal of Science and Technology*, 9: 1-7. https://doi.org/10.17485/ijst/2016/v9i48/106000.

[24] McCallum A., Nigam K., Ungar L. H. (2000, August). Efficient clustering of high-dimensional data sets with application to reference matching. *In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 169-178. https://doi.org/10.1145/347090.347123.

[25] Kumar A., Ingle Y. S., Pande A., Dhule P. (2014). Canopy clustering: a review on pre-clustering approach to K-Means clustering. *Int. J. Innov. Adv. Comput. Sci.(IJIACS)*, 3(5): 22-29.

[26] Ali T., Asghar S., Sajid N. A. (2010, June). Critical analysis of DBSCAN variations. *In 2010 International Conference on Information and Emerging Technologies, IEEE,* pp. 1-6. https://doi.org/ 10.1109/ICIET.2010.5625720.

[27] Ester M., Kriegel H. P., Sander J., Xu X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. *In Kdd*, 96(34): 226-231.

[28] Moreira A., Santos M. Y., Carneiro S. (2005). Density-based clustering algorithms–DBSCAN and SNN. *University of Minho-Portuga*l.

[29] Chakraborty S., Nagwani N. K. (2011). Analysis and study of Incremental DBSCAN clustering algorithm. *International Journal of Enterprise Computing and Business Systems*, 1: 101-130. https://doi.org/10.1007/978-3-642-22577-2_46.

[30] Gu X., Angelov P. P., Kangin D., Principe J. C. (2017). A new type of distance metric and its use for clustering. *Evolving Systems*, 8(3): 167-177. https://doi.org/10.1007/s12530-017-9195-7.

[31] Vijaykumar V. R., Jothibasu P. (2010, September). Decision based adaptive median filter to remove blotches, scratches, streaks, stripes and impulse noise in images. *In 2010 IEEE International Conference on Image Processing, IEEE,* pp. 117-120. https://doi.org/10.1109/ICIP.2010.5651915.

[32] Chen T., Wu H. R. (2001). Adaptive impulse detection using center-weighted median filters. *IEEE signal processing letters*, 8(1): 1-3. https://doi.org/10.1109/97.889633.

[33] Zhao Y., Li D., Li Z. (2007, August). Performance enhancement and analysis of an adaptive median filter. *In 2007 Second International Conference on Communications and Networking in China, IEEE,* pp. 651-653. https://doi.org/10.1109/CHINACOM.2007.4469475

[34] Sasaki Y. (2007). The truth of the F-measure. *Teach Tutor mater*, 1(5): 1-5. https://doi.org/10.13140/RG.2.1.1571.5369.

[35] Gokilam G. G., Shanthi K. (2015). Comparing clustering Algorithms with Diabetic Datasets in WEKA Tool.

https://doi.org/10.31449/inf.v43i4.2461 Informatica **43** (2019) 507–513 **507**

An Adaptive Image Inpainting Method Based on the Weighted Mean

Nguyen Hoang Hai
Faculty of Information Technology, University of Science and Education, The University of Danang, Vietnam
E-mail: hoanghai@ued.udn.vn

Le Minh Hieu
Department of Economics, University of Economics, The University of Danang, Vietnam
E-mail: hieulm@due.udn.vn

Dang N. H. Thanh,
Department of Information Technology, Hue College of Industry, Vietnam
E-mail: dnhthanh@hueic.edu.vn

Nguyen Van Son
Ballistic Research Laboratory, Military Weapon Institute, Vietnam
E-mail: vanson.nguyen.mwi@gmail.com

V. B. Surya Prasath
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, USA
Department of Pediatrics, University of Cincinnati, USA
Department of Biomedical Informatics, College of Medicine, University of Cincinnati, USA
Department of Electrical Engineering and Computer Science, University of Cincinnati, USA
E-mail: surya.prasath@cchmc.org, prasatsa@uc.edu

Keywords: inpainting, weighted mean, weighted mean filter, image restoration, image processing

Received: September 26, 2018

Imaging inpainting is the process of digitally filling-in missing pixel values in images and requires carefully crafted image analysis tools. In this work, we propose an adaptive image inpainting method based on the weighted mean. The weighted mean is assessed to be better than the median because, for the case of the weighted mean, we can exclude the values of the corrupted pixels from evaluating values to fill those corrupted pixels. In the experiments, we implement the algorithm on an open dataset with various corrupted masks and we also compare the inpainting result by the proposed method to other similar inpainting methods – the harmonic inpainting method and the inpainting by directional median filters – to prove its own effectiveness to restore small, medium as well as fairly large corrupted regions. This comparison will be handled based on two of the most popular image quality assessment error metrics, such as the peak signal to noise ratio, and structural similarity. Further, since the proposed inpainting method is non-iterative, it is suitable for implementations to process big imagery that traditionally require higher computational costs, such as the large, high-resolution images or video sequences.

Povzetek: Opisana je novo razvita metoda določanja manjkajočih točk v sliki s pomočjo uteženega povprečja.

1 Introduction

Image inpainting [1, 2, 3, 4, 5, 6] is one of the interesting problems of image processing [7, 8] that has attracted a lot of attention. Image inpainting or image interpolation is a process of filling the damaged and/or missing parts. Image inpainting has a wide range of applications in practice such as watermark removal [9], image disocclusion [10, 11], restoring old images corrupted by dust, scratches, etc., image zooming, image super-resolution [8], etc.

To solve the image inpainting problem, there are some intensive approaches [12, 13, 14] such as partial-differential-equation-based (PDE-based) methods [1, 8, 15], calculus-of-variation-based methods [8], graph-based methods, stochastic methods, etc. and machine-learning-based methods. In this paper, we only focus on non-learning-based methods, because the comparison of methods based on machine-learning to non-learning-based methods is not fair. In non-learning-based methods, the PDE-based and variation-based methods are highly effective to treat this problem. However, these methods are iterative and the accuracy, performance, execution time depends much on the tolerance.

In this paper, we study an inpainting method based on the weighted mean [16]. Our method is non-iterative, hence, the computational complexity associated with traditional iterative inpainting methods. This advantage is especially appealing as it paves the way for us to apply

inpainting for larger images or high-resolution images, such as Full HD 1080, 2K, 4K, and higher resolution and/or video sequences data.

We propose the method to fill the corrupted parts by the weighted mean because it excludes the value of the corrupted pixels from the evaluating process. The pixels of the corrupted regions are always different from the neighboring pixels' values. So, their values are not useful for evaluating the values to fill the corrupted regions. This idea is better than using the median. The proposed inpainting method is effective on various corrupted masks: from a small, medium up to large area.

In our experiments, we compare our inpainting results from the proposed method to the harmonic inpainting method [1, 15] (PDE-based inpainting) and the inpainting method by directional median filters [17]. Our results prove that the adaptive inpainting based on the weighted mean is a novel method and performs favorably to other state-of-the-art inpainting methods.

The rest of the paper is organized as follows. Section 2 introduces the general image inpainting problem that is formulated in the form of the optimization problem, and the proposed adaptive inpainting method based on the weighted mean. Section 3 presents the experiments and the comparison of the inpainting result to another similar method. Finally, section 4 concludes.

2 The proposed inpainting method

2.1 Image inpainting problem

Let $u_0(x,y), u(x,y), v(x,y), (x,y) \in \Omega \subset \mathbb{R}^2$, be the 2D grayscale original image, the restored image, and the corrupted image, respectively. In the case of image presented in the discrete form, $x = 1,2 \dots, m; y = 1,2, \dots, n$, where values m, n are the number of pixels by the image width and image height. Let $\mathcal{D}$ – the set of corrupted pixels.

The popular general image inpainting model is defined as follows [1, 15]:

$$u = \underset{u}{\operatorname{argmin}} \left(\sum_{x,y \in \Omega} \Phi(\nabla u) + \lambda \sum_{(x,y) \in \Omega \setminus \mathcal{D}} |u - v|^2 \right),$$

where the norm $|\cdot|$ – L^2 norm, $\Phi(\cdot)$ – a gradient-based smooth function. It can be selected as Total variation [18, 19], Euler's elastica [1] or Mumford-Shah model [1].

As can be observed from the minimization model, to solve the above inpainting model, we need some complex optimization methods. The algorithms based on this model are iterative manners that the accuracy depends much on the tolerance.

2.2 The proposed inpainting method

In this paper, we propose the inpainting method based on the adaptive weighted mean filter. The adaptive weighted mean filter [16] is proposed by Zhang and Li to solve the denoising problem. Like other median-based and mean-based filters, this method evaluates the corrupted pixels values on the clipping windows by the mean value. However, this method considers the weight of pixels values to evaluate the mean. So, it is called to be the weighted mean.

For the inpainting problem, the corrupted regions are the connected regions with a small, medium or large area. This is quite different from the denoising problem: the corrupted pixels are always separated pixels or connected pixels on a small area (for the high noise levels).

Like the adaptive weighted mean filter for denoising, we also evaluate the corrupted pixels values based on the pixels values of the considered clipping windows.

Let I, I_c be indices of pixels of all image domain Ω and of the corrupted regions $\mathcal{D}$, respectively. We call $W_{ij}(d)$ a clipping window centered at the pixel (i,j) with a size of $2d+1$:

$$W_{ij}(d) = \{|i-k| \le d, |j-l| \le d, (k,l) \in I\}.$$

Figure 1 shows an example of the clipping windows centered at the pixel (3, 3) with $d = 2$.

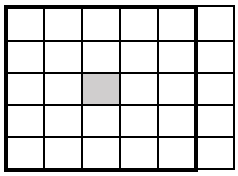


Figure 1: The clipping window centered at the marked pixel (3, 3) with a $d = 2$ (i.e., the window size is 5).

u_{22}	u_{21}	u_{21}	u_{22}	u_{23}	u_{23}	u_{22}
u_{12}	u_{11}	u_{11}	u_{12}	u_{13}	u_{13}	u_{12}
u_{12}	u_{11}	u_{11}	u_{12}	u_{13}	u_{13}	u_{12}
u_{22}	u_{21}	u_{21}	u_{22}	u_{23}	u_{23}	u_{22}
u_{22}	u_{21}	u_{21}	u_{22}	u_{23}	u_{23}	u_{22}
u_{12}	u_{11}	u_{11}	u_{12}	u_{13}	u_{13}	u_{12}

Figure 2: The extended pixels matrix with the padding is 2 of the marked 3×2 image.

Our proposed inpainting method starts from the clipping windows of every corrupted pixel $(i,j) \in I_c$. If the corrupted pixels lay on the position (i,j) with $i - d \le 0$ or $i + d \ge m$ or $j - d \le 0$ or $j + d \ge n$, we need to extend the pixels matrix to top, bottom, left or right more d rows or columns of pixels. The values to fill to the padding regions are taken by the symmetric method. In MATLAB, we can use the built-in function *padarray*. Figure 2 shows an example of the padding pixels matrix of the image 3×2 (the gray pixels area) with the padding is 2.

Next step, we need to evaluate the maximum, the minimum and the weighted mean values in every clipping window centered at the corrupted pixels:

$$A = \max\{W_{ij}(d)\}, (i,j) \in I_c,$$

$$B = \min\{W_{ij}(d)\}, (i,j) \in I_c,$$

$$\mu = m\ddot{e}an\{W_{ij}(d)\} = \begin{cases} \sum_{(k,l)\in W_{ij}(d)} a_{k,l}v_{kl}, if \sum_i a_{k,l} \neq 0 \\ -1, \text{ otherwise} \end{cases}, (i,j) \in I_c,$$

where

$$a_{k,l} = \begin{cases} 1, & if \quad \min\{W_{ij}(d)\} < v_{k,l} < \max\{W_{ij}(d)\} \\ 0, & otherwise \end{cases}.$$

Finally, with every pixel of the corrupted regions $(i,j) \in I_c$, we replace the corrupted pixel value by the weighted mean value μ.

The proposed image inpainting method has a smaller number of calculations than the denoising method by the adaptive weighted mean filter because we only consider the clipping windows of the corrupted pixels. The algorithm of the proposed image inpainting method by the adaptive weighted mean is presented in Algorithm 1.

Algorithm 1. The inpainting method base on the adaptive weighted mean.

Input: The corrupted image v.
Output: The restored image u.

Initialize $h = 1, d_0 = 1, d_{max} = 39$.
For every pixel $(i,j) \in I_c$
  Set $d = d_0$.
  While $(d \leq d_{max})$
    Evaluate $A = \max\{W_{ij}(d)\}$.
    Evaluate $A' = \max\{W_{ij}(d+h)\}$.
    Evaluate $B = \min\{W_{ij}(d)\}$.
    Evaluate $B' = \min\{W_{ij}(d+h)\}$.
    Evaluate $\mu = m\ddot{e}an\{W_{ij}(d)\}$.
    If ($A == A'$ and $B == B'$ and $\mu \neq -1$) **Then**
      Set $u_{ij} = \mu$.
      Break.
    Else
      Increase $d = d + h$.
      If $(d > d_{max})$ **Then**
        Set $u_{ij} = \mu$.
        Break.
      End.
    End.
  End.
End.

As can be seen in Algorithm 1, the values of the corrupted pixels are replaced by the weighted mean value. The weighted mean value will not contain the values of corrupted pixels in the clipping windows. We choose the weighted mean value because the values of the corrupted pixels always have no similarity with the values of other neighboring pixels in the same clipping window.

3 Experiments

We implement the proposed inpainting method and the harmonic inpainting method to recover the corrupted image by MATLAB 2018a. The configuration of the computing system is Windows 10 Pro with Intel Core i5, 1.6GHz, 4GB 2295MHz DDR3 RAM memory. For the harmonic inpainting method, we set the tolerance $Tol = 10^{-5}$. This value will balance the accuracy and execution time.

3.1 Image quality assessment metrics

To compare the inpainting result of the proposed method to other similar inpainting methods, it is necessary to assess image quality after inpainting based on the error metrics. The popular error metrics are PSNR and SSIM that were used in many works [20, 21, 22, 23, 24, 25, 26, 27]:

$$PSNR = 10\log_{10}\left(\frac{u_{max}^2}{MSE}\right) dB,$$

where

$$MSE = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(u^{(ij)} - u_0^{(ij)}\right)^2$$

is the mean squared error with u_0 is the original (latent) image, u_{max} denotes the maximum value, for e.g. for 8-bit images $u_{max} = 255$; $u^{(ij)}$ and $u_0^{(ij)}$ are pixels values of u and u_0 at every pixel (i,j). The difference of 0.5dB is visible. The higher PSNR (measured in decibels – dB), the better image quality.

Structural similarity (SSIM) is proven to be a better error metric for comparing the image quality and it is in the range [0, 1] with a value closer to one indicating better structure preservation. This metric based on the characteristic of the human vision. The SSIM is computed between two windows and of common size $N \times N$,

$$SSIM = \frac{(2\mu_{\omega_1}\mu_{\omega_2} + c_1)(2\sigma_{\omega_1\omega_2} + c_2)}{(\mu_{\omega_1}^2 + \mu_{\omega_2}^2 + c_1)(\sigma_{\omega_1}^2 + \sigma_{\omega_2}^2 + c_2)},$$

where μ_{ω_i} – the average of ω_i, $\sigma_{\omega_i}^2$ – the variance of ω_i, $\sigma_{\omega_1\omega_2}$ – the covariance, and c_1, c_2 numerical stabilizing parameters.

IDs	Corrupted	Harmonic method	Directional median	Proposed
119082	16.9766	25.2881	25.5074	**25.9331**
126007	17.3561	30.1535	31.4514	**31.6833**
157055	19.248	26.962	27.4127	**28.3869**
170057	18.3731	31.0244	32.2728	**33.1103**
182053	18.5361	25.8607	28.2344	**28.5674**
219090	17.2918	28.1766	30.4586	**30.7027**
253027	18.3034	24.7708	25.3588	**25.7218**
295087	17.013	28.6833	32.2062	**32.2184**
296007	19.4597	33.2096	36.2832	**36.3026**
38092	19.2699	27.1973	29.597	**29.7045**

Table 1: The comparison of PSNR metric of the inpainting methods for the mask 1.

3.2 Image datasets and test cases

We test the performance of the proposed inpainting method on a dataset of natural images with artificial corrupted masks. The dataset is well-known as BSDS of UC Berkeley *https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/BSDS300/html/dataset/images.html*.

All images are stored in JPEG format, grayscale and have size 481×321 pixels. Figure 3 shows the selected image of UC Berkeley for the tests. The corrupted masks are presented in Figure 4. In this case, the black areas indicate the corrupted regions. Note that, we use 10 original images and 2 masks. Hence, our experiment will be handled on 20 images. This size is suitable for non-learning-based methods without any trained data.

IDs	Corrupted	Harmonic method	Directional median	Proposed
119082	0.90632	*0.85151*	0.90542	**0.91662**
126007	0.88987	0.89776	0.94538	**0.95302**
157055	0.91292	*0.88074*	0.92816	**0.9373**
170057	0.89567	0.90925	0.95351	**0.95409**
182053	0.90673	*0.87299*	0.9434	**0.94673**
219090	0.89305	*0.86256*	0.94373	**0.94649**
253027	0.90216	*0.84878*	0.93184	**0.9351**
295087	0.89063	*0.83746*	0.94324	**0.94702**
296007	0.89367	*0.89022*	0.94741	**0.95609**
38092	0.91695	*0.83039*	0.93749	**0.93903**

Table 2: The comparison of SSIM metric of the inpainting methods for the mask 1.

IDs	Corrupted	Harmonic method	Directional median	Proposed
119082	13.1181	24.7802	24.9254	**25.0446**
126007	14.0723	28.5263	28.5795	**28.8802**
157055	17.7535	27.2226	27.9019	**28.376**
170057	14.5046	29.8771	30.5236	**31.0405**
182053	14.6563	24.9808	26.1866	**26.2575**
219090	13.7254	26.5841	26.9604	**27.5309**
253027	15.252	23.1356	23.1589	**23.3926**
295087	14.0301	27.7979	29.6814	**29.7502**
296007	15.372	32.887	35.0773	**35.158**
38092	14.0467	26.1847	27.4583	**27.5423**

Table 3: The comparison of PSNR metric of the inpainting methods for the mask 2.

In the case of the mask 1, the corrupted ratio is 5%. The size of the corrupted parts (white square) is 14×14 pixels. The inpainting results are presented in Figure 5. The values of the PSNR metric are presented in table 1. The value of the SSIM metric is shown in table 2. We can see that the inpainting result by the harmonic method is good by the PSNR metric, but it fails for SSIM. The SSIM of the harmonic inpainting method is smaller than one of the corrupted images. However, in Figure 5, the quality of the inpainting result by the harmonic is really better than the corrupted image. The harmonic method changes whole the image (including the uncorrupted regions). The restored image gets smoother. This is the reason to make SSIM metric of the harmonic inpainting method to be lower. The inpainting result by the directional median filters is better, but there are still some defects: flowers on the dress of woman (ID 157055), on leaves and branches of the tree (ID 295087).

IDs	Corrupted	Harmonic method	Directional median	Proposed
119082	0.80231	0.82849	0.88415	**0.88549**
126007	0.78606	0.87759	0.9203	**0.92621**
157055	0.82274	0.87131	0.91147	**0.91767**
170057	0.77925	0.89494	0.93326	**0.93476**
182053	0.81001	0.85041	0.90769	**0.91723**
219090	0.78722	0.83843	0.90378	**0.91129**
253027	0.8193	0.81883	0.88859	**0.89715**
295087	0.78451	0.81901	0.90972	**0.91991**
296007	0.77249	0.87662	0.93102	**0.93571**
38092	0.80371	0.80615	0.90268	**0.90704**

Table 4: The comparison of SSIM metric of the inpainting methods for the mask 2.

Figure 3: The input natural images from the BSDS dataset of UC Berkeley.

For our proposed method, the inpainting result is very good. Our method only recovers the value of the corrupted pixels. Because the corrupted area is not large, our proposed method gives perfect inpainting result. By both PSNR and SSIM metrics, our proposed method is better than the harmonic method and directional median filters.

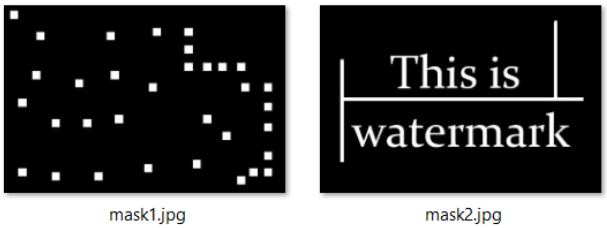


Figure 4: The masks for generating the corrupted regions.

In the case of the mask 2, the corrupted ratio is 10%. The lengths of some corrupted regions are very long. The corrupted areas, in this case, are bigger than one of the cases of the mask 1. The inpainting results are presented in Figure 6. The PSNR and SSIM metrics are presented in Table 3 and Table 4, respectively. The harmonic method still made the whole image to be smoother. The inpainting result by the directional median filters looks better. The inpainting result of our proposed method is best. Our proposed method preserves the structure of the image because it did not change the pixels' values outside the corrupted regions. Both PSNR and SSIM metrics of the proposed method, in this case, are also better than ones of the harmonic method and the directional median filters.

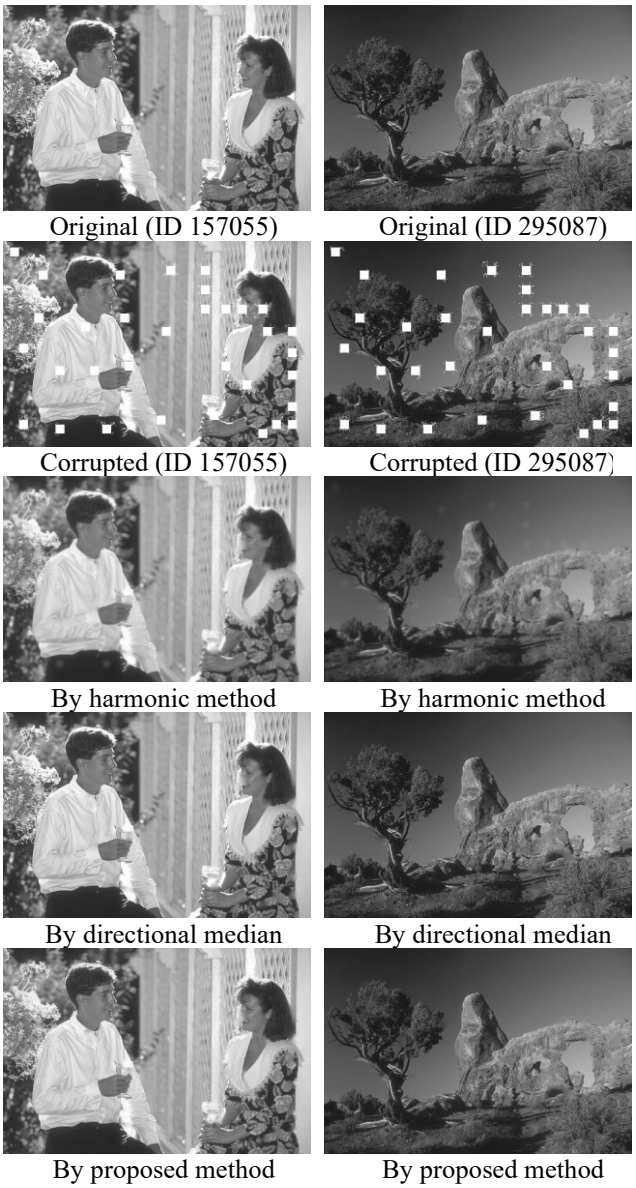


Figure 5: The inpainting results on the corrupted image with the mask 1.

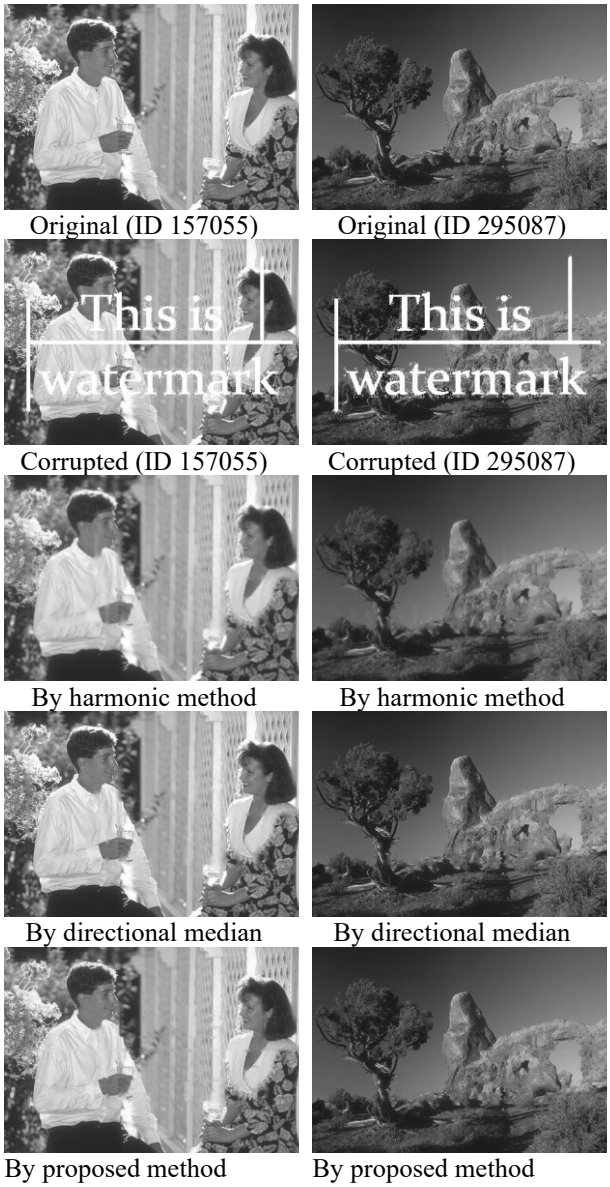


Figure 6: The inpainting results on the corrupted images with the mask 2.

From the above experiments, we can see that the proposed method can work well on various masks with small and medium corrupted regions. For the large corrupted regions, our proposed method also works well. However, if the corrupted area is too large, our method will work ineffectively, because the proposed inpainting method uses the clipping windows to evaluate. If a lot of pixels in the clipping windows are corrupted, the recovery process cannot get high accuracy. Figure 7 shows the inpainting results of the proposed method on the large corrupted regions (the corrupted ratio is 42%, the corrupted sizes are 30 pixels by width and 30 pixels by height). As can be seen, our proposed method creates a "paintbrush" effect. By human eyes, the inpainting result is good, and the values of PSNR and SSIM metrics of the inpainting result for this case by the proposed inpainting method are good enough too PSNR=21.9475, SSIM=0.7029.

We have to notice that, in real applications of image inpainting, the most important task is to detect the mask. The mask is not only the corrupted regions on images but also is objects that need to be removed. This task depends on every problem. It can be made manually or automatically. For example, in order to segment skin lesions of dermoscopic images [28], we need to detect hairs and remove them to improve image quality. The hairs can be detected by many methods, including machine learning methods, curvatures-based techniques, etc.

As we discussed above, our proposed method is a non-iterative manner, and it only considers the clipping windows of the corrupted pixels, so it can work very fast. In all the tests, the execution time is under 1 second. The harmonic inpainting method spent up to 40 seconds with above tolerance 10^{-5}. This is promising as the method is suitable to process large, high-resolution images or video sequences that are our future applications.

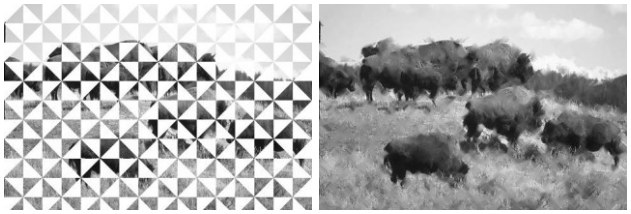


Figure 7: The inpainting results by the proposed method on the large corrupted regions.

4 Conclusion

In this paper, we proposed an adaptive inpainting method based on the weighted mean. The proposed method can restore the images with small and medium corrupted regions effectively. It only recovers the corrupted pixels and do not touch other pixels. Hence, the image structure is unchanged. For the large corrupted areas, our proposed method works well, but it creates an artificial effect – the paintbrush effect.

The proposed inpainting method is non-iterative manner, so it can work very fast. Otherwise, the proposed method only considers the clipping windows of every corrupted pixels, so it requires less memory. These advantages are useful to process large, or high-resolution images or video sequences.

In future works, we would like to improve the proposed method to remove the paintbrush effect during restoring the images with large corrupted areas.

5 References

[1] C. B. Schönlieb, Partial Differential Equation Methods for Image Inpainting, Cambridge: Cambridge University Press, 2015.

[2] H. Grossauer, Digital Image Inpainting: Completion of Images with Missing Data Regions, Innsbruck: Simon & Schuster, 2008.

[3] D. N. H. Thanh, V. B. S. Prasath, N. V. Son and L. M. Hieu, "An Adaptive Image Inpainting Method Based on the Modified Mumford-Shah Model and Multiscale parameter estimation," Computer Optics, vol. 43, no. 2, pp. 251-257, 2019. http://doi.org/10.18287/2412-6179-2019-43-2-251-257

[4] D. N. H. Thanh, V. B. S. Prasath, L. M. Hieu and K. Hiroharu, "Image Inpainting Method Based on Mixed Median," in 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Spokane, 2019. http://doi.org/10.1109/ICIEV.2019.8858556

[5] U. Erkan, S. Enginoglu and D. N. H. Thanh, "An Iterative Image Inpainting Method Based on Similarity of Pixels Values," in IEEE 6th International Conference on Electrical and Electronics Engineering (ICEEE), Istanbul, 2019. http://doi.org/10.1109/ICEEE2019.2019.00028

[6] D. N. H. Thanh, N. V. Son and V. B. S. Prasath, "Distorted Image Reconstruction Method with Trimmed Median," in IEEE 3rd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom), Hanoi, 2019. http://doi.org/10.1109/SIGTELCOM.2019.8696138

[7] D. N. H. Thanh, V. B. S. Prasath, S. Dvoenko, L. M. Hieu, "An Adaptive Image Inpainting Method Based on Euler's Elastica with Adaptive Parameters Estimation and the Discrete Gradient Method," Signal Processing, 2020 (In press).

[8] T. F. Chan and J. Shen, Image Processing and Analysis: Variational, PDE, Wavelet and Stochastic Methods, SIAM, 2005.

[9] L. Li, Y. Y. Ma, C. C. Chang and J. F. Lu, "Analyzing and removing SureSign watermark," Signal Processing, vol. 93, no. 5, pp. 1374-1378, 2013. http://doi.org/10.1016/j.sigpro.2012.10.001

[10] S. Masnou and J. M. Morel, "Level-lines based disocclusion," in 5th IEEE International conference on Image processing, 3:249-263, Chicago, 1998. http://doi.org/10.1109/ICIP.1998.999016

[11] Z. Tauber, Z. N. Li and M. S. Drew, "Review and Preview: Disocclusion by Inpainting for Image-Based Rendering," IEEE Transactions on Systems, Man, and Cybernetics, vol. 37, no. 4, pp. 527-540, 2007. http://doi.org/10.1109/TSMCC.2006.886967

[12] J. Cheng and Z. Li, "Markov random field-based image inpainting with direction structure distribution analysis for maintaining structure coherence," Signal Processing, vol. 154, pp. 182-197, 2019. http://doi.org/10.1016/j.sigpro.2018.09.004

[13] D. Ding, S. Ram and J. J. Rodríguez, "Image Inpainting Using Nonlocal Texture Matching and Nonlinear Filtering," IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 1705 - 1719, 2019. http://doi.org/10.1109/TIP.2018.2880681

[14] H. Liu, X. Bi, G. Lu and W. Wang, "Exemplar-Based Image Inpainting With Multi-Resolution Information and the Graph Cut Technique," IEEE Access, vol. 7, pp. 101641 - 101657, 2019. http://doi.org/10.1109/ACCESS.2019.2931064

[15] J. Shen and T. F. Chan, "Mathematical models for local nontexture inpaintings," SIAM Journal on Applied Mathematics, vol. 62, no. 3, pp. 1019-1043, 2002. http://doi.org/10.1137/S0036139900368844

[16] P. Zhang and F. Li, "A New Adaptive Weighted Mean Filter for Removing Salt-and-Pepper Noise," IEEE Signal Processing Letters, vol. 21, no. 10, p. 1283, 2014. http://doi.org/10.1109/LSP.2014.2333012

[17] H. Noori and S. Saryazdi, "Image Inpainting Using Directional Median Filters," in IEEE International Conference on Computational Intelligence and Communication Networks, Bhopal, 2010. http://doi.org/10.1109/CICN.2010.20

[18] D. N. H. Thanh and S. Dvoenko, "A method of total variation to remove the mixed Poisson-Gaussian noise," Pattern Recognition and Image Analysis, vol. 26, no. 2, pp. 285-293, 2016. http://doi.org/10.1134/S1054661816020231

[19] D. N. H. Thanh and S. Dvoenko, "Image noise removal based on Total Variation," Computer Optics, vol. 39, no. 4, pp. 564-571, 2015. http://doi.org/10.18287/0134-2452-2015-39-4-564-571

[20] V. B. S. Prasath, D. N. H. Thanh and N. H. Hai, "On Selecting the Appropriate Scale in Image Selective Smoothing by Nonlinear Diffusion," in IEEE 7th International Conference of Communications and Electronics, pp. 267-272, Hue, 2018. http://doi.org/10.1109/CCE.2018.8465764

[21] D. N. H. Thanh, S. Dvoenko, and D. V. Sang, "A Denoising Method Based on Total Variation," in Proceedings of the Sixth International Symposium on Information and Communication Technology (SoICT 2015), pp. 223-230, Hue, 2015. http://doi.org/10.1109/10.1145/2833258.2833281

[22] Z. Wang, A. Bovik, H. Sheikh, Simoncelli and Eero, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004. http://doi.org/10.1109/TIP.2003.819861

[23] U. Erkan, D. N. H. Thanh, L. M. Hieu and S. Enginoglu, "An Iterative Mean Filter for Image Denoising," IEEE Access, vol. 7, no. 1, pp. 167847-167859, 2019. http://doi.org/10.1109/ACCESS.2019.2953924

[24] D. N. H. Thanh, V. B. S. Prasath, L. M. Hieu and S. Dvoenko, "An Adaptive Method for Image Restoration Based on High Order Total Variation and Inverse Gradient," Signal, Image and Video Processing, 2020 (In press).

[25] D. N. H. Thanh, L. T. Thanh, N. N. Hien and V. B. S. Prasath, "Adaptive Total Variation L1 Regularization for Salt and Pepper Image Denoising," Optik - International Journal for Light and Electron Optics, 2019 (In press). http://doi.org/10.1016/j.ijleo.2019.163677

[26] P. Jidesh and H. K. Shivarama, "Non-local total variation regularization models for image restoration," Computers & Electrical Engineering, vol. 67, pp. 114-133, 2018. http://doi.org/10.1016/j.compeleceng.2018.03.014

[27] V. B. S. Prasath and D. N. H. Thanh, "Structure tensor adaptive total variation for image restoration," Turkish Journal Of Electrical Engineering & Computer Sciences, vol. 27, no. 2, pp. 1147-1156, 2019. http://doi.org/10.3906/elk-1802-76

[28] D. N. H. Thanh, V. B. S. Prasath, L. M. Hieu and N. N. Hien, "Melanoma Skin Cancer Detection Method Based on Adaptive Principal Curvature, Colour Normalisation and Feature Extraction with the ABCD Rule," Journal of Digital Imaging, 2019 (In press). http://doi.org/10.1007/s10278-019-00316-x

https://doi.org/10.31449/inf.v43i4.2598 Informatica **43** (2019) 515–525 **515**

Determination of Blood Flow Characteristics in Eye Vessels in Video Sequence

Chaoxiang Chen, Shiping Ye and Huafeng Chen
Zhejiang Shuren University, 8, Shuren Str., 310015, Hangzhou, China
E-mail: eric.hf.chen@hotmail.com

Alexander Nedzvedz and Sergey Ablameyko
Belarusian State University, 4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus
United Institute of Informatics Problems of National Academy of Sciences
6, Surganova str., 220020, Minsk, Republic of Belarus

Olga Nedzvedz
Belarusian State Medical University, 83, Dzerzhinski Ave., 220116, Minsk, Republic of Belarus

Keywords: blood flow, eye conjunctiva, vascular net, integral optical flow

Received: November 27, 2018

Accurately measuring blood flow in eye is an important challenge, as blood flow reflects the health of eye and is disrupted in many diseases. Existing techniques for measuring blood flow are limited due to the complex assumptions and calculations required. Digital image and video processing techniques started to be used for eye vessels analysis and evaluation during last decades. In this paper, we propose a method for determining the characteristics of blood flow in the vessels of eye conjunctiva, such as linear and volumetric blood speed, and topological characteristics of vascular net. The method first analyses image frame by frame sequentially and then builds integral optical flow for video sequence. Dynamic characteristics of eye vessels are introduced and calculated. These characteristics make it possible to determine changes in blood flow in eye vessels. We show the efficiency of our method in real eye vessels scenes.

Povzetek: Razvit je nov sistem za določanje pretoka krvi v očeh.

1 Introduction

The study of conjunctiva vessels allows to perform a direct non-invasive study of vessels of a microcirculatory bed. The change in the quantitative dynamic characteristics of blood flow in the conjunctival vessels determines the change in blood flow in a microvascular bed and reaction of a vascular bed to the effects of various therapeutic drugs. It appears in narrowing or dilation of blood vessels, increasing degree of branching, the expansion of a capillary network.

Currently, there are many methods for blood flow monitoring, such as Doppler ultrasonography and Velocimetry, laser Doppler flowmetry, and blood vessels measurement by portable devices [1,2]. Volumetric blood flow speed is usually computed as the product of linear blood speed and vessel cross-sectional area, each measured independently. Vessel diameter can be measured from conjunctiva images by using the retinal vessel analyzer or the scanning laser ophthalmoscope, or by using confocal line scans. Speed can be measured with bidirectional laser Doppler velocimetry and frequency domain optical coherence tomography [3].

However, most of the methods allow to determine the parameters of blood flow only in straight sections of the microvascular bed, while the changes occurring in the nodes and complex fragments of the vascular network remain unaccounted for. In addition, the use of these methods in clinical studies is limited due to the high cost and complexity of result interpretation.

The image of eye circulatory system is a network of vessels of different shapes, sizes, orientations and brightness. The location of vessels allows to get a fairly clear image without gross distortion. However, in the study of these images there are some difficulties, some of them are typical for vessels of any part of a circulatory system, others - only for vessels of eye circulatory system.

The first problem that arises when obtaining a video sequence is caused by image instability. For a healthy human, eye is characterized by saccadic eye movements, which are rapid jumps of different duration and amplitude from one point of fixation to another, as well as eye tremor of different intensity, depending on the state of human health. When receiving such images, it is impossible to fix an object. Thus, mixing of vessel positions between two frames is chaotic. Therefore, at this stage, the most important task is to stabilize a video sequence.

The second problem is to determine the structure of conjunctiva circulatory system. It is a structure with complex geometry, which is characterized by a large number of vessels with bends and branches. The vessel is a complex three-dimensional structure and we have to take

into account changes in shape, size and brightness. Branching and intersecting vessels also complicate the task of segmentation.

The third problem is related to changes in the shape and geometric characteristics of vessels due to the fact that vessel is an elastic object, and its geometrical parameters change with blood filling during a cardiovascular cycle. To reduce the measurement error, a vessel diameter should be determined at the time of the lowest blood filling of the vessel. Also, in addition to the average geometry parameters and blood flow volumetric speed, it is necessary to take into account the vessel instantaneous characteristics.

The fourth problem is the complexity of describing sequence of events occurring in the vessel due to periodic changes in the parameters describing the blood flow in conjunctival vessels.

To obtain images of the conjunctiva, a monochrome camera equipped with a laser device for guidance and focus is used. Despite the use of pulse illumination synchronization devices, the obtained images have low quality and resolution, which leads to the need to use complex methods of image analysis.

Most of the existed methods process eye vessels in single static images. However, blood flow characteristics can be only computed by using sequence of images or video sequence and special techniques should be elaborated for this task.

In this paper, we propose a method for determining the characteristics of blood flow in the vessels of eye conjunctiva, such as linear and volumetric blood speed, topological characteristics of vascular net. The method first analyses image frame by frame sequentially and then builds integral optical flow for video sequence. Dynamic characteristics of eye vessels are introduced and calculated. These characteristics make it possible to determine changes in blood flow in eye vessels. We show the efficiency of our method in real eye vessels scenes.

2 Review of eye vessels image analysis approaches

Image and video processing techniques started to be used for eye vessels analysis during last decades. It allows to organize an efficient permanent control of eye state and define the corresponding treatment. On the other hand, it requires advanced techniques for image and video processing.

Methods of eye vessels image analysis are considered in many papers. The first reviews of algorithms for processing of vascular structures can be seen in [4, 5]. The review [4] presented an analysis and categorization of literature related to digital imaging technologies in the field of diabetic retinopathy and focused on algorithms and methods of segmentation of two-dimensional color retina images that are received with the help of fundus cameras.

Authors in paper [6] presented a comparative overview of methods and algorithms for isolating vessels and elongated objects on both two-dimensional and three-dimensional medical images used in various tasks. A review of the algorithms for segmentation and registration of the retina is presented in paper [7], that is mainly devoted to tasks of detecting boundaries and central lines of the vessels. Paper [8] provides an overview of algorithms primarily focused on the isolation of vessels on two-dimensional color images of retina obtained with fundus cameras or fluorescent angiography, and focused on studies related to segmentation of blood vessels of retina.

From recent results, we can indicate paper [9], where authors propose a novel contextual method for analysis of vessel connectivities based on the geometry of the primary visual cortex. Using the spectral clustering on a large local affinity matrix constructed by both the connectivity kernel and the feature of intensity, the vessels are identified successfully in a hierarchical topology each representing an individual perceptual unit. Paper [10] presents an algorithm for segmenting and measuring retinal vessels, by growing an active contour model, which uses two pairs of contours to capture each vessel edge, while maintaining width consistency. The algorithm is initialized using a generalized morphological order filter to identify approximate vessels centerlines. Once the vessel segments are identified, the network topology is determined using an implicit neural cost function to resolve junction configurations. Paper [11] proposes several methods for vessels image segmentation. One method uses Matched Filter (MF) for the extraction of blood vessels. This method responds not only to vessels but also to non-vessel edges. The second method is a novel hybrid automatic approach for extraction of retinal image vessels which reduce the weak edges and noise, and finally extract the blood vessels.

Vessels tracing is considered in many papers. Paper [12] proposes a novel graph-based approach to address this tracing with crossover problem. After initial steps of segmentation and skeleton extraction, its graph representation can be established, where each segment in the skeleton map becomes a node, and a direct contact between two adjacent segments is translated to an undirected edge of the two corresponding nodes. The segments in the skeleton map touching the optical disk area are considered as root nodes. This determines the number of trees to-be-found in the vessel network, which is always equal to the number of root nodes.

An automatic algorithm capable of segmenting the whole vessel tree and calculate vessel diameter and orientation in a digital ophthalmologic image is presented in paper [13]. The algorithm is based on a parametric model of a vessel that can assume arbitrarily complex shape and a simple measure of match that quantifies how well the vessel model matches a given angiographic image.

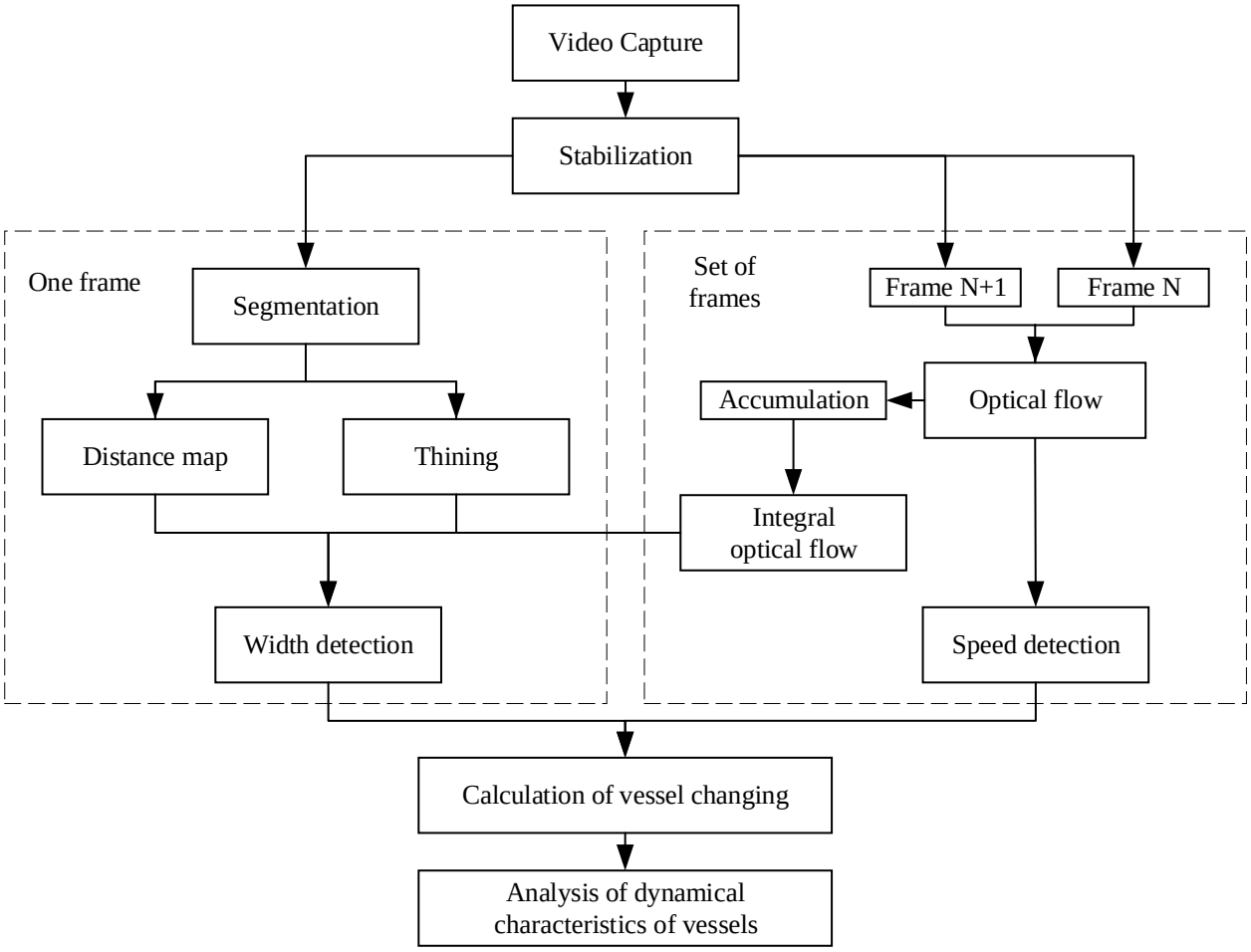


Figure 1: Scheme of video sequence processing and definition of blood flow speed in vessel.

An automated screening system to diagnose the retinal images affected by diabetic retinopathy is described in paper [14]. The proposed system consists of three stages: the pre-processing, which is done to make the image reliable for extracting features. In the second stage, features like area of blood vessels and texture features were extracted from the retinal images and classification – the last stage was done using the ELM classifier. The very recent comprehensive overview for retinal vessels segmentation techniques is given in paper [15].

Most of the existed methods can segment and process retina cells in single static images, they are important for eye vessels analysis. However, blood flow characteristics can be only computed by using sequence of images or video sequence.

3 Video sequence capture and processing

A general scheme of video sequence processing and definition of blood flow speed in vessel is shown in Figure 1. Video sequence processing can be divided into several parts. Video capture constructs an image sequence for further processing. The first step of video processing is stabilization.

Video stabilization is performed in the following way. The first frame in video sequence plays an important role in the whole process and it should be prepared in a specific way. The contrast of vessels in an image varies and is often quite low, so first it is necessary to manually define a fragment with a clear image of the vessel to search stable regions for it in subsequent frames. This search is performed through the analysis of field borders using the bitmap filter, for example, Sobel filter.

The search for the maximum brightness to determine the contrast begins at the center of the frame, because objects in the image can move in any direction in the next frame. The region of interest is defined as the sharpest fragment located as close to the center of image as possible. According to new positions of the selected fragment on the subsequent frames, the offsets relative to the first frame are calculated and intermediate images are created.

Intermediate images are used as the core of correlation to determine the coordinate offsets when stabilizing the video. Then, based on the calculation of the image shifts, video sequence frames are aligned and video sequence is stabilized. Stabilization ensures constant positions of vessels in each frame, which allows monitoring at each given coordinate.

The next processing block consist of two branches. The first branch is segmentation of vessels at image frame. The second is going for a limited set of neighboring images from sequence for flow speed determination. The

set of images is used for smoothing of characteristics of flow motion.

4 Frame image segmentation by using neural networks

4.1 Image preprocessing and accumulation

Image preprocessing is used first to adjust image brightness, correct irregularities, suppress noise, and eliminate distortion. To do this, standard operations are used, for example, histogram equalization to increase the contrast of vascular images. Segmentation allows to select certain fragments in the image of the network of blood vessels. Then, numerical data on blood flow in the areas allocated during segmentation are estimated. The obtained data can be used to classify objects according to predefined criteria such as size, structure, and brightness.

The next step is frame image segmentation, which allows to select fragments of blood vessels in the image. Blood vessels are objects, segmentation of which is quite difficult. This is due to numerous vascular occlusions, complex bends and branches, and variabilities of size and brightness of objects in the images. Furthermore, in video sequence, vessels are without blood filling in some frames, which makes the vessels invisible in these frames.

For improving the quality of segmentation, the structure of vessels needs to be determined. Segmentation is performed on a synthesized image that corresponds to the normalized integral sum of all video sequence frames. To obtain it, the accumulation of images is performed.

The accumulated image is then used to improve the image of vessels before segmentation. This is done by averaging this image with the current image frame:

$$I = \frac{1}{n}\sum_{k=0}^{n}(I\,(n-1) + I_n),$$

where I is the brightness of the intermediate image; n is the number of frames that have already been processed; I_n is the current image.

The advantage of the synthesized image is the absence of fragments of blood vessels, not filled with blood. As a result, all vessels became visible and thus it is possible to highlight them.

4.2 Segmentation by convolutional neural network

In every day medical practice, image segmentation is usually performed manually by doctors, which is time-consuming and tedious. However, the ever-increasing quantity and variety of medical images make manual segmentation impracticable in terms of cost and reproducibility. Therefore, automatic medical image segmentation is highly desirable. However, images with vessels of eye conjunctiva are very complex due to complex variations in objects and structures and because of low contrast, noise, and other imaging artifacts caused by various imaging modalities and techniques.

Automatic image segmentation methods started to widely be used in the last few years for medical images. These methods achieve promising results on nonmalignant objects using hand-crafted features and prior knowledge of structures. However, the automatic segmentation of eye vessels images does not give desirable result and usually interactive (manual) postprocessing should be used.

Recently neural networks, particularly fully convolutional networks, have been proved highly effective for medical image segmentation, which require little hand-crafted features or prior knowledge. Neural networks might be able to take the position, size, shape, intensity, etc. and do a better job of figuring out where the required objects are in an image compared to simply applying morphological operations or other segmentation methods. With strong use of data augmentation, this segmentation model achieves significant improvement over previous methods.

The method of teaching a convolutional neural network (CNN) with a sliding window was used in our study. This technique makes it possible to predict the class label for each pixel, based on the pattern selected around it, according to [16,17]. That is, a small area around the pixel is used as the source data.

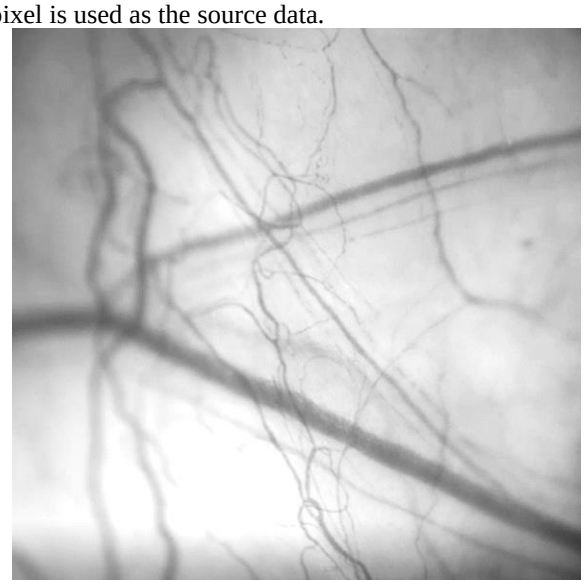
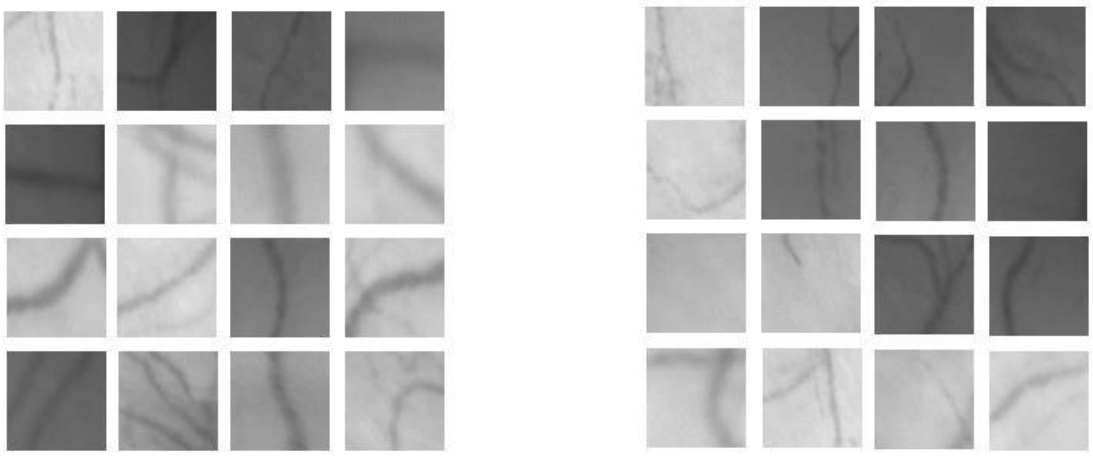


Figure 2: Image of vessels at eye sclera.

To solve the problem of vascular network segmentation, a fully connected convolutional neural network was used. The peculiarity of its organization is that the usual convolution network is supplemented with layers in which the union operators are replaced by operators of increasing discretization, which leads to an increase in the resolution of the output image. Combining features with higher resolution from a narrowing area with an expanding output area allows to train convolutional layers to form a more accurate result at the output.

A set of 130 gray-scale images of eye sclera using a GigE camera was used to train the neural network (Figure 2).



a) b)

Figure 3: Example of training patterns: a) with vessel in the center (class 0), b) without vessel in the center (class 1).

The SNA architecture proposed during the isbi 2012 EM Segmentation Challenge (Segment Neural Membranes) was used [17]. Neural network training was performed on a set of images, which was increased by simple geometric augmentation to 650 synthetic images.

We have chosen the following transformation for augmentation: 1) flips, 2) turns, 3) reflections, 4) elastic deformations, and 5) scaling.

The training took place on the NVIDIA GTX GPU. The training lasted for 500 epochs. All regions containing vessel in the center and other regions without vessel in the center with the same size were selected from each of the input sample images (Figure 3). The optimization method for CNN is a stochastic gradient descent. The network output is the probability for each vessel from 0 to 1, where 0 stands for vessel, 1 stands for non-vessel.

When the amount of data increases, neural network is retrained to detect vessels. This means that it is necessary to submit more teaching images, without a vessel in the center. It is especially important to include images where the center pixel lies near the vessel, but does not belong to it, so that the network learns to identify borders of the vessel. At the same time, it is important to allow mixing of any subset of the training data. The number of objects of class 0 and 1 must be identical.

Let us define number of all patterns containing the vessel in the center is V, and number of all patterns without the vessel in the center is NV. Then, in the teaching process, only N patterns from those NV patterns are randomly determined. It allows to use a larger set of data without increasing the load on the neural network and avoid network retraining. On this basis, the neural network becomes more flexible. As a result, the quality of segmentation has increased.

The network architecture is shown in Figure 4. It consists of two almost symmetrical parts: narrowing left and expanding right.

The narrowing part corresponds to the typical architecture of SNA and consists of two sequentially applied convolution blocks of size 3x3 (no indent). Each block is a ReLU layer and the operation of subsampling (2x2 max pooling) with step 2 decreases the image size. After each decrease in the dimension, the number of features is doubled. Every step of the growing branch consists of research layers (convolution 2x2 to increase the resolution) and formed on their basis a set of attributes, which reduces by half the number of signs.

The neural network has 23 convolutional layers, to bring each 64-component vector to the required number of classes, convolutions of 1x1 size are applied on the last layer. The size of the input image is determined by the need for even values of height and width for the proper operation of subsampling (2x2 max pooling).

Then, there is a concatenation with the corresponding set of features from the narrowing part, and two 3x3 convolutions are performed, each of which is followed by a transformation through the ReLU activation function [16]. The neural network with this architecture has demonstrated good results for the segmentation of blood vessels in ophthalmology.

Usually this neural network requires a long teaching time, but it can be compensated by the high segmentation rate of the trained network. Full HD (1920 x 1080) image resolution on the NVIDIA GTX 950 GPU takes less than 10 seconds to fully segment, which is acceptable for medical image processing.

4.3 Segmentation results

As a result of segmentation based on convolutional neural networks, the image of the vascular network is obtained (Figure 5).

We used standard deviation and accuracy mark for estimation of effectiveness of this algorithm [5]. Standard deviation is calculated as:

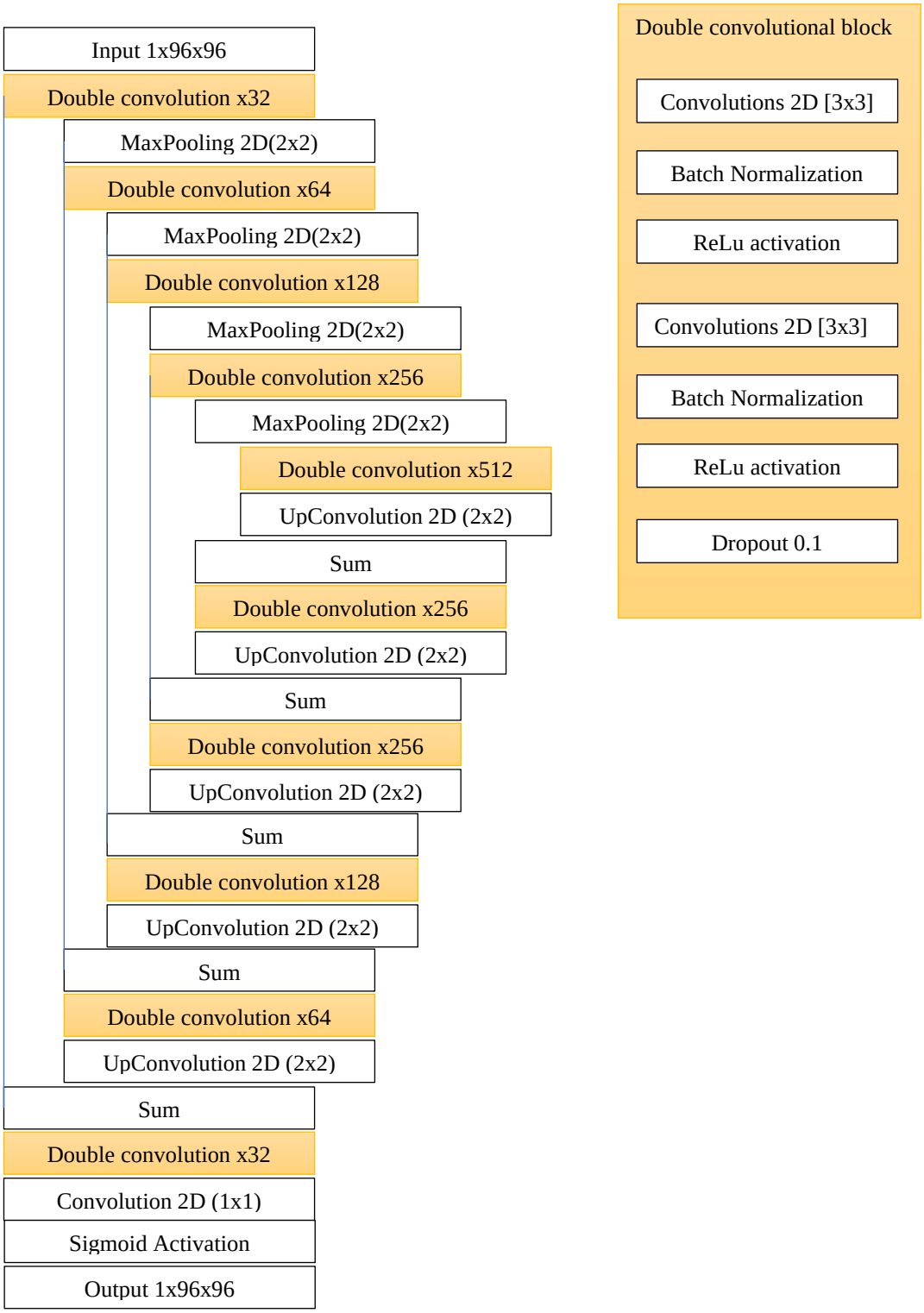
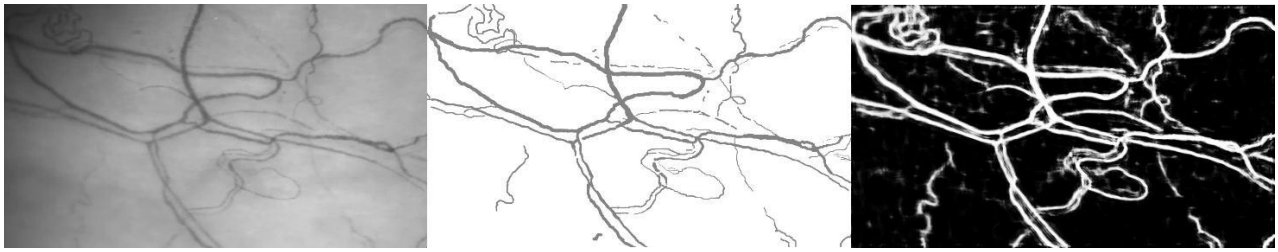


Figure 4: U-net architecture of used neural network.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where n is pixel number, x_i is result label (0 or 1), $\bar{x}$ – probability result. Accuracy is normalized number of true answers. Estimation of segmentation effectiveness is shown in Table 1. As it is shown, the CCN-based segmentation algorithm allows to extract regions of vessels from gray-scale image with high quality. Our experiments proved that convolutional neural networks can be successfully used to segment such complex images as images of vessels at eye sclera.



a) b) c)

Figure 5: Segmentation of vessels image: a) source vessels image, b) interactive segmentation by user, c) result of segmentation by CNN.

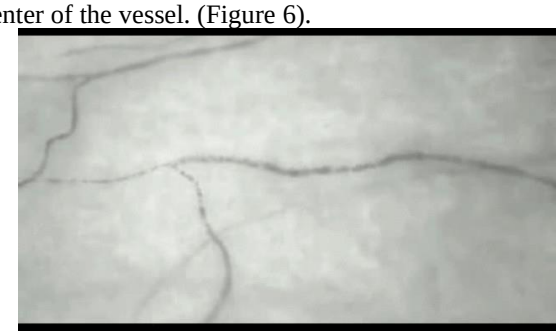
Data	Accuracy	Standard deviation
Vessels detection	0.9415	0.1997
Regions without vessels	0.9201	0.2035
Common mean estimation	0.9308	0.2016

Table 1: Estimation of effectiveness of the segmentation.

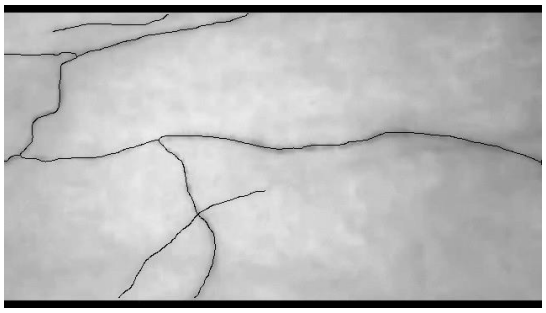
5 Blood flow characteristics calculation by using optical flow

Segmentation still has errors and blood flow speed has all sorts of boundary effects, so we try to avoid this by limiting area of objects. In this case, the optimal vessel region for analysis is a middle line of a vessel that is obtained by thinning operation. After segmentation is performed, the instantaneous linear speed at each point of vessel is determined. To do this, the method based on optical flow is used. Therefore, when preparing the image of vascular network for calculation of optical flow, its thinning is first performed.

For analysis of blood flow speed, the optical flow is determined only for the part of the image directly containing vessels. Resulting video sequence displays all the changes occurring in vessels. Such a transformation can significantly reduce the contribution of events occurring outside the vessels. The calculation of the optical flow for points along the midline of the capillary allows to analyze the instantaneous linear speed in the center of the vessel. (Figure 6).



a)



b)

Figure 6: Image of vessels a) source image, b) image with a skeleton after thinning.

5.1 Optical flow determination

For description convenience, we use I_t to denote t-th frame of video I and $I_t(p)$ to denote pixel with coordinate $p = (x, y)$ in I_t.

Let OF_t denote basic optical flow of I_t. It is a vector field with each vector $OF_t(p)$ represents displacement vector of pixel $I_t(p)$. Assume $OF_t(p) = \vec{d}$, we can easily determine the coordinate in I_{t+1} where pixel $I_t(p)$ moves, and it is $p + \vec{d}$.

Considering optical flows for several consecutive frames have been computed, we can obtain integral optical flow for the first frame of those. Let IOF_t^{itv} denote integral optical flow of I_t, where itv is the frame interval parameter used to compute integral optical flow [18]. IOF_t^{itv} is also a vector field which records accumulated displacement information in time period of itv frames for all pixels in I_t.

For any pixel $I_t(p)$, its integral optical flow $IOF_t^{itv}(p)$ can be determined as follows:

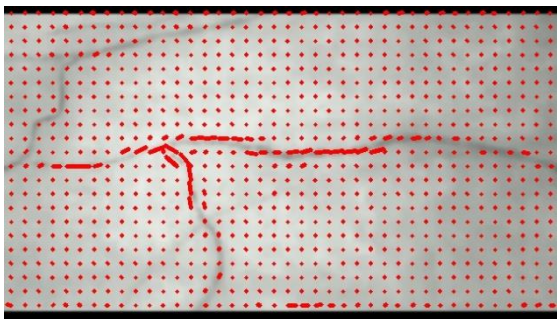
$$IOF_t^{itv}(p) = \sum_{i=0}^{itv-1} OF_{t+i}(p_{t+i}),$$

where p_{t+i} is the coordinate in I_{t+i} of pixel $I_t(p)$. In other words, if $I_t(p)$ stays in the video scene, $I_t(p_t), I_{t+1}(p_{t+1}), \cdots, I_{t+itv-1}(p_{t+itv-1})$ are the same pixel in different frames, i.e. $I_t(p)$. Note that x-component and y-component of p_{t+i} should be rounded to the nearest integer, as pixels are at coordinates with integer values.

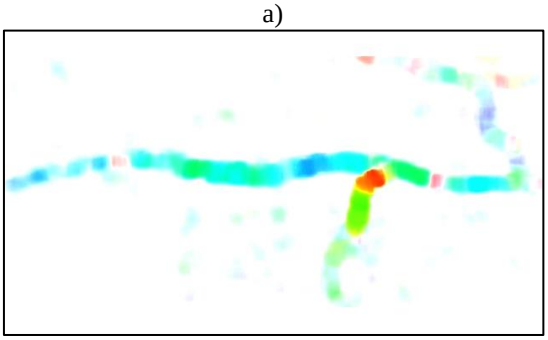
Optical flow indicates the speed of blood flow through the vessel. At almost every point of the vessel skeleton, we have an instant speed of blood flow.

This method allows us to estimate the displacement at each point of the image between two frames of the video sequence and is based on the determination of the intensity shift for a short period of time. After the segmentation process, the images are processed using a binary mask of the vascular network. This operation is performed on the basis of masking with skeleton of vascular network with images of each frame of video sequence. This makes it possible not to take into account the change in brightness in the vicinity of the vessel when calculating the optical flow throughout the image.

5.2 Blood flow speed calculation by using optical flow



a)



b)

Figure 7: Optical flow field: a) vector representation, b) color representation.

After vessels image thinning, analysis of its midline neighborhoods is performed, and the optical flow is calculated. To compute the optical flow, the algorithm [19] was used. As a result, an array of vectors for the vertical and horizontal speed components are calculated (Figure 7a). With the help of the polar transformation, amplitudes and directions of these vectors are determined (Figure 7b).

Then, a new image, in which intensity corresponds to the magnitude and hue corresponds to the direction of optical flow vector, is created. To determine the speed of blood flow, only the magnitude (amplitude) is used. Thus, one can build a profile of the speed values along the midline of the skeleton (Figure 8).

Figure 8: Masking optical flow and skeleton.

This profile represents a change in blood instantaneous linear speed for any point on the middle line of the vessel (Figure 9).

Problems related to the discretization of time and space make it difficult to use absolute values. The optical flow values were used to determine the instantaneous linear speed, which was measured in relative units. The volumetric speed of blood flow in the capillary depends on its width, it can be calculated by the formula:

$$Q = v \cdot A_v,$$

where v is the linear speed of blood flow, Q is the volumetric speed, A_v is the cross-section area of the vessel.

A cross-section area of the vessel is calculated from width vessel as:

$$A_v = \frac{\pi d^2}{4},$$

where d is diameter of vessel in separate point.

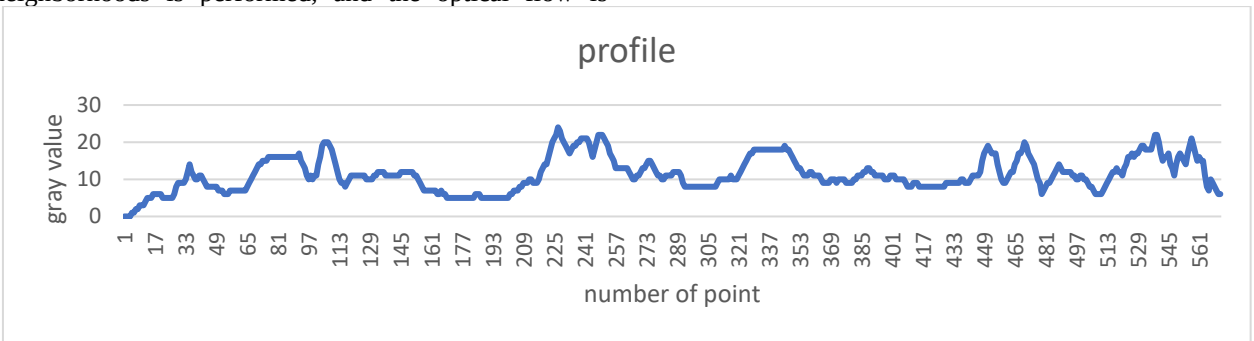


Figure 9: The intensity profile for the line of a skeleton reflecting change of linear speed of a blood flow.

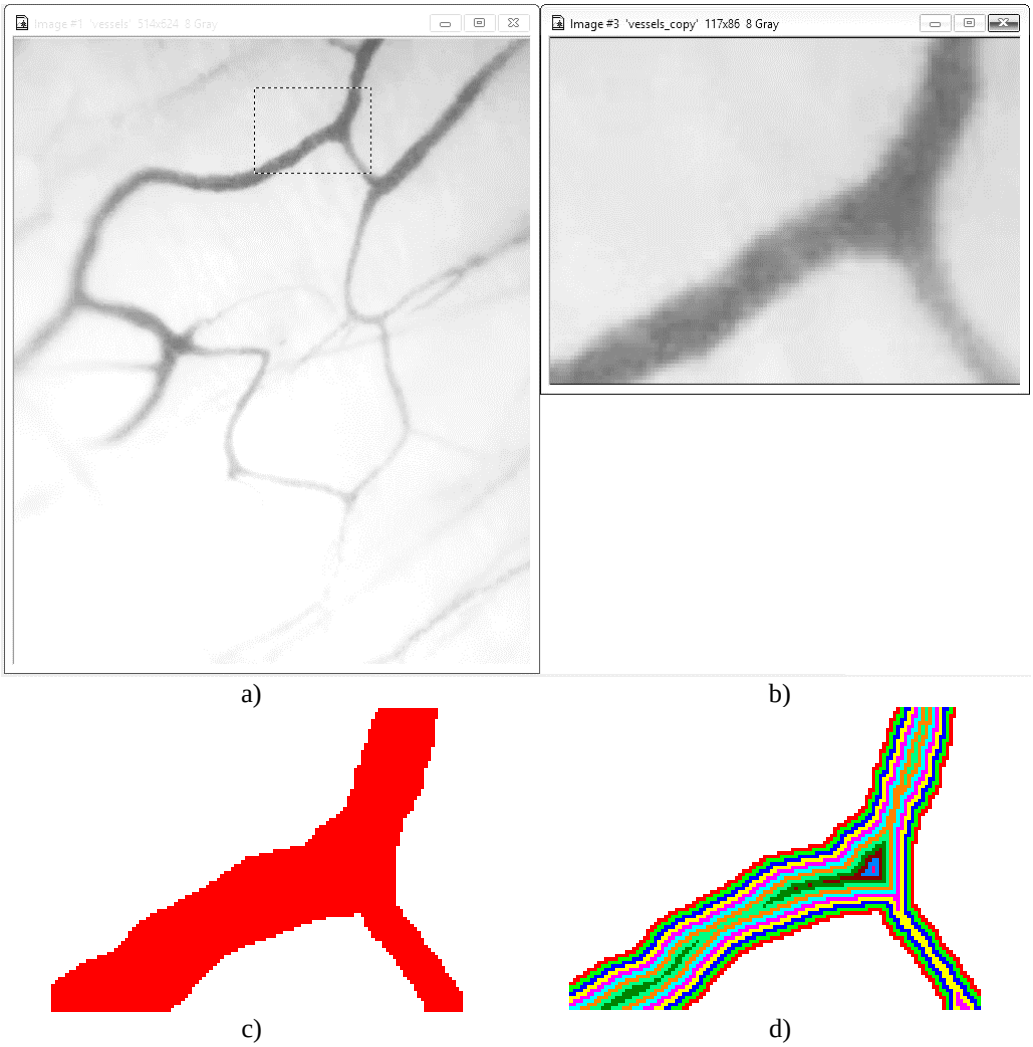


Figure 10: Diameter (d) calculation by using distance map: a) initial image, b) fragment of vessel with diameter marker, c) binarization result, d) distance map and diameter calculation.

Diameter corresponds to width of vessel. Determining the width of the vessel is a complex task due to unstable diameter of a blood vessel section. Currently, there is no algorithm for qualitative construction of the distribution of vascular width. In this study, we determine a distribution of width based on distance map (Figure 10) analysis. It allows to determine the width change along vessels.

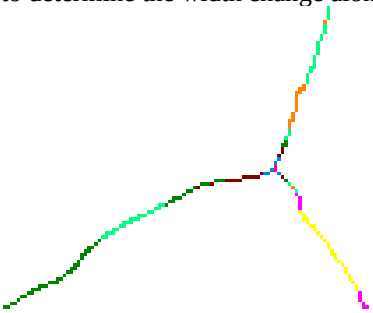


Figure 11: Vessel skeleton with width values defined for each point.

Distribution of the width along vessels is constructed on base of the distance map ridges. This operation is based on the construction of the labeled skeleton (Figure 11), in which color indicates width of the vessel.

It allows to get important practical information about the features of the blood flow through the vascular network, taking into account its geometric complexity. The speed of blood flow in a vessel and its diameter are used to determine the instantaneous changes occurring in the vessel. Additionally, determination of speed based on optical flow and vessel width can be performed in parallel, which help reduce computational cost.

5.3 Determination of topological characteristics of vascular net

Vessel segments are characterized by width, length and blood flow speed. Diameter and speed are obtained by averaging results of multiple adjacent blood vessel speed profiles. Other hemodynamic parameters can be calculated by topological description.

On the base of above-described automatic segmentation and morphological identification, the skeleton of the conjunctiva vascular net is detected, and branch points of the vessels are automatically indicated. The following parameters are calculated for vascular net: length, branchiness, compactness, and tortuosity, based on

determination of the skeleton structure including nodes, segments and tails (Figure 12).

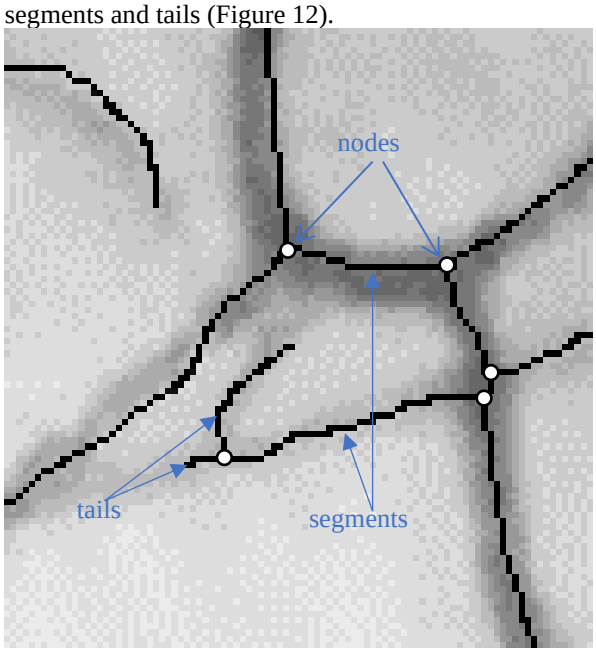


Figure 12: Vessel skeleton with the values of width defined for each point.

It is known that in some pathologies the tortuosity of eye vessels increases. Analysis of the topology of the vascular net is performed on the crucial elements of the skeleton.

We define *nodes* as branching nodes, *skeleton* as vascular net, *segments* as vessel sections between branching points, *area(vessel)* as area of vessel, *count()* as function of quantity determination, and *length()* as function of length determination, respectively.

Branchiness and *curliness* describe complexity of vessels net. They are defined by ratio of node number to length of skeleton and to count of skeleton segments correspondingly. The skeleton segment is fragment of skeleton between nodes or skeleton ends.

$$branchiness = \frac{count(nodes)}{length(skeleton)},$$

$$curliness = \frac{count(nodes)}{count(segments)}.$$

Definition of tails allows to produce characteristics for description of topological properties with complexity of vascular net. They are tailness, tails curliness, and tails ratio.

$$tailness = \frac{count(tails)}{length(skeleton)},$$

$$tails\ curliness = \frac{count\ (tails)}{count(segments)}.$$

Full length of vascular net corresponds to length of skeletons.

$$vessel\ length = length(skeleton),$$

$$mean\ vessel\ width = \frac{area(vessel)}{length(skeleton)}.$$

These characteristics and their combination with flow speed allow to describe hydrodynamics properties of vascular net. It can be used in monitoring and diagnostics. Comparing the data obtained before and after treatment, the clinicians receive objective information about the newly formed vessels and other changes in the vascular net.

6 Discussion and conclusion

In this paper, we propose a method for determining the characteristics of blood flow in the vessels of eye conjunctiva, such as linear and volumetric blood speed, and topological characteristics of vascular net. The method first analyses image frame by frame sequentially and then builds integral optical flow for video sequence. Dynamic characteristics of eye vessels are introduced and calculated. These characteristics make it possible to determine changes in blood flow in eye vessels. We show the efficiency of our method in real eye vessels scenes.

The method was tested on the video sequence of blood vessels of conjunctiva. The change in blood flow speed in vessels reflects the change in blood flow in the microcirculatory bed, as well as in various organs in normal and pathological conditions. The study was conducted using a high-resolution monochrome digital video camera Imperx Bobcat IGV-B1410M with a microscope lens with a focal length of 40 mm.

The linear speed of blood flow in a vessel with a diameter of 1.91 μm is 0.50379 relative units, which corresponds to 5·10-5 m / s. This result is consistent with the data obtained by the Doppler method.

The proposed method is designed to study the characteristics of the blood-vascular net. It is based on the determination of the instantaneous linear and volumetric speed for each point of the vessel. The method allows to carry out a quantitative assessment of the cross-section area, and linear and volumetric speed in the vessels in normal and in various pathologies.

We define the following characteristics of vessels: branchiness, curliness, tailness, tails curliness, tails ratio, vessel length, and mean vessel width. These topological and dynamical characteristics allow to detect new possibilities for eye vessels analysis during healing process. That allows to quantify changes in the linear speed of blood flow in the vessels of healthy people in the simulation of hypercapnia and hyperoxia.

In comparison with the known methods based on static images analysis, our method allows to detect and study blood flow in eye vessels in dynamics. Such description allows to predict effectiveness of treatment.

The defined characteristics make it possible to determine changes in blood flow in the microcirculatory bed, which in turn determine changes in blood flow in the vessels of the brain, kidneys, and coronary vessels.

Acknowledgement

This work is supported by Public Welfare Technology Applied Research Program of Zhejiang Province (LGJ19F020002, LGJ18F020001 and LGF19F020016) and the National High-end Foreign Experts Program (GDW20183300463).

References

[1] C. J. Pournaras and C. E. Riva. Retinal blood flow evaluation. *Ophthalmologica*, 229(2): 61-74, 2013. https://doi.org/10.1159/000338186

[2] T. E. Kornfield and E. A. Newman. Measurement of Retinal Blood Flow Using Fluorescently Labeled Red Blood Cells. *eNeuro*, 2(2):1-13, 2015. https://doi.org/10.1523/ENEURO.0005-15.2015

[3] P. Ganesan, S. He, and H. Xu. Analysis of retinal circulation using an image-based network model of retinal vasculature. *Microvascular Research*, 80(1): 99-109, 2010. https://doi.org/10.1016/j.mvr.2010.02.005

[4] R. J. Winder, P. J. Morrow, I. N. McRitchie, J. R. Bailie, and P. M. Hart. Algorithms for digital image processing in diabetic retinopathy. *Computerized Medical Imaging and Graphics*, 33(8): 608-622, 2009. https://doi.org/10.1016/j.compmedimag.2009.06.003

[5] K. Bühler, P. Felkel, and A. L. Cruz. Geometric methods for vessel visualization and quantification - a Survey. *In: Brunnett G., Hamann B., Müller H., Linsen L. (eds) Geometric Modeling for Scientific Visualization. Mathematics and Visualization. Springer, Berlin, Heidelberg*, pp. 399-421, 2004. https://doi.org/10.1007/978-3-662-07443-5_24

[6] C. Kirbas and F. Quek. A review of vessel extraction techniques and algorithms. *ACM Computing Surveys*, 36(2): 81-121, 2004. https://doi.org/10.1145/1031120.1031121

[7] M. S. Mabrouk, N. H. Solouma, and Y. M. Kadah. Survey of retinal image segmentation and registration. *International Journal on Graphics, Vision and Image Processing*, 6(2): 1-11, 2006.

[8] A. R. Rudnicka, C. G. Owen, and S. A. Barman. Blood vessel segmentation methodologies in retinal images. *Computer Methods and Programs in Biomedicine*, 108(1): 407-433, 2012. https://doi.org/10.1016/j.cmpb.2012.03.009

[9] M. Favali, S. Abbasi-Sureshjani, B. H. Romeny, and A. Sarti. Analysis of Vessel Connectivities in Retinal Images by Cortically Inspired Spectral Clustering. *Journal of Mathematical Imaging and Vision*, 56(1): 158-172, 2016. https://doi.org/10.1007/s10851-016-0640-1

[10] B. Al-Diri, A. Hunter, and D. Steel. An active contour model for segmenting and measuring retinal vessels. *IEEE Transactions on Medical Imaging*, 28(9): 1488-1497, 2009. https://doi.org/10.1109/TMI.2009.2017941

[11] H. S. Bhadauria1, S. S. Bisht, and A. Singh. Vessels Extraction from Retinal Images. *IOSR Journal of Electronics and Communication Engineering*, 6(3): 79-82, 2013. https://doi.org/10.9790/2834-0637982

[12] J. De, H. Li, and L. Cheng. Tracing retinal vessel trees by transductive inference. *BMC Bioinformatics*, 15: 20, 2014. https://doi.org/10.1186/1471-2105-15-20

[13] K. K. Delibasis, A. I. Kechriniotis, C. Tsonos and N. Assimakis. Automatic model-based tracing algorithm for vessel segmentation and diameter estimation. *Computer Methods and Programs in Biomedicine*, 100(2): 108-122, 2010. https://doi.org/10.1016/j.cmpb.2010.03.004

[14] I. S. H. Punithavathi and P. G. Kumar. Extraction of Blood Vessels for Retinal Image Analysis. *Middle-East Journal of Scientific Research*, 24(1): 450-457, 2016.

[15] J. Almotiri, K. Elleithy, and A. Elleithy. Retinal Vessels Segmentation Techniques and Algorithms: A Survey. *Applied Sciences*, 8(2): 155, 2018. https://doi.org/10.3390/app8020155

[16] A. Nedzvedz, O. Nedzvedz, A. Glinsky, G. Karapetian, I. Gurevich, and V. Yashina. Detection of dynamical properties of flow in an eye vessels by video sequences analysis. *In Proceeding of International Conference on Information and Digital Technologies*, IEEE, Zilina, pp. 275-281, 2017. https://doi.org/10.1109/DT.2017.8024308

[17] D. Ciresan, D. A. Giusti, L. M. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. *In Proceeding of Advances in Neural Information Processing Systems*, pp. 2843-2851, 2012.

[18] C. Chen, S. Ye, H. Chen, O. V. Nedzvedz, and S. V. Ablameyko. Integral optical flow and its application for monitoring dynamic objects from a video sequence. *Journal of Applied Spectroscopy*, 84(1): 120-128, 2017. https://doi.org/10.1007/s10812-017-0437-z

[19] G. Farneback. Two-frame motion estimation based on polynomial expansion. *In Proceedings of the 13th Scandinavian Conference on Image Analysis*, Springer, Halmstad, pp. 363-370, 2003. https://doi.org/10.1007/3-540-45103-X_50

Refin-Align: New Refinement Algorithm for Multiple Sequence Alignment

Ahmed Mokaddem, Amine Bel Hadj and Mourad Elloumi
Laboratory of Technologies of Information and Communication, and Electrical Engineering, University of Tunis, Tunisia
E-mail:moka.ahmed@yahoo.fr, amine_bel_hadj_90@yahoo.com, Mourad.Elloumi@gmail.com

Keywords: multiple sequence alignment, algorithms, refinement, block

Received: December 17, 2018

In this paper, we present Refin-Align a new refinement algorithm for a multiple sequence alignment. Refining alignment consists on constructing a new more accurate multiple sequence alignment from an initial one by applying some modifications. Our refinement algorithm Refin-Align uses a new definition of block and also our multiple sequence alignment algorithm Pro-malign. We assess our algorithm Refin-Align on multiple sequence alignment constructed by different algorithms using different benchmarks of protein sequences. In the most cases treated, our algorithm improves the scores of the multiple sequence alignment.

Povzetek: Različni znani algoritmi napovedujejo zaporedje beljakovin, novo razviti algoritem pa določa najboljšo skupno vrednost na osnovi napovedi posameznih algoritmov.

1 Introduction

Multiple sequence alignment is an important task in bioinformatics. Aligning a set of sequences consists in optimizing the number of matches between the characters occurring in the same order in each sequence (figure1).

w_1 :W T Y I - M R E A Q Y E S A Q
w_2 :W T C I V M R E A - Y E - - -
w_3 :W - Y I - M Q E V Q Q E R - -
w_4 :W R Y I A M R E - Q Y E S - -
w_5 :W - Y I A - R E - Q Y E S - -
w_6 :W T - I A M R E - Q Y E S - -

Figure 1: Multiple sequence alignment.

Multiple sequence alignment can help biologist to predict structure and function information for a set of sequences. Indeed, we can reveal information about biological functions common to biological macromolecules from several different organisms by identifying similar regions, these regions are often an important structural or functional roles. Multiple sequence alignment can also help in the classification of macromolecules into different families according to similar sub-strings detected. In addition, multiple sequence alignment can help to construct a phylogenetic tree and analyse relationships between species in order to establish a common biological ancestor.

Although pairwise sequence alignment for two sequences can be constructed with optimal solution using the dynamic programming algorithm [1], multiple sequence alignment for more than two sequences is a NP-complete problem [2]. There are two main approaches to resolve this problem:

1. Progressive approach: it consist to align sequences gradually. Indeed, we start by aligning the most similar two sequences. Then, we align the sequences to other sequences aligned, according to a defined order. Finally, we obtain the multiple sequence alignment. All progressive multiple sequence alignment algorithms adopt the same process. The most used progressive multiple sequence algorithms are ClustalW [3], T-COFFEE [4], MUSCLE [5], MAFFT [6], GLProbs [7] and Clustal Omega [8].

   Progressive approach operates in three steps:

   (a) In the first step, we compute distances between all pairs of sequences of the set and we store these distances in a matrix called *distance matrix*. This step aims to estimate the similarity between pairs of sequences in order to distinguish the two sequences that are the first to be aligned. Many distances are used [9]. Among these distances we mention:
   - *k-mer distances* used by the algorithm MUSCLE and MAFFT,
   - *Percent of similarity* used by the algorithm ClustalW,
   - *Kimura distance* [10] used by the algorithm Clustal Omega,
   - Distance defined by the GLProbs algorithm.

   (b) In the second step, we construct a guide tree using the distance Matrix. This step aims to define the order of aligning sequences. Two main algorithms are used to construct a guide tree:
   - UPGMA [11] used by MUSCLE, MAFFT and GLProbs

- Neighbor-joining [12] used by ClustalW and T-COFFEE

(c) In the last step, we follow the branching order of the guide tree, constructed in the previous step, to construct the multiple sequence alignment by aligning pair of sequences using the dynamic programming algorithm[1] or by *a profile-profile*[3] alignment.

A *profile* is constructed by selecting for each column of the sequence alignment the character that have the maximum occurrences in that column (Figure 2).

W_1: W - Y I - M Q E V Q Q E R
W_2: W R Y I A M R E - Q Y E S
profile: W - Y I - M - E - Q - E -

Figure 2: Profile construction.

2. Iterative approach: it consists to construct an initial multiple sequence alignment. Then, we apply a number of iterations, during each iteration we perform a set of modifications to the current alignment in order to ameliorate his score. Among this modifications, we can insert or delete of one or more gaps '-' in one or more position in the multiple of sequence alignment. The main multiple sequence alignment algorithms adopting iterative approach are genetic algorithm such as GAPAM [13] and PASA [14].

Each algorithm adopting progressive approach or iterative approach produces mistakes in multiple sequence alignment, thus, we used refinement algorithms in order to correct bad aligned residues, that can ameliorate the quality of the multiple alignment by ameliorate his scores. The process of all refinement algorithms consists to apply a set of modifications to an initial multiple sequence alignment in order to construct a new one having better scores than the previous alignment. These modifications are repeated until *convergence* (i.e. no improvement can be made on the current alignment). There are different algorithms for refinement of multiple sequence alignments:

1. RASCAL [15]: Rascal operates as follows: First, we analyse the initial multiple sequence alignment and detect the well-aligned regions by applying the Mean Distance (MD). Then, we detect the badly aligned regions. Finally, we realign the badly aligned regions.

2. REFINER [16]: when applying REFINER algorithm on a multiple sequence alignment, we realign each sequence with the profile of the multiple sequence alignment of the remaining sequences. Convergence is obtained when all the iterations is realised and each sequence is realigned.

3. RF [17]: is similar to the REFINER algorithm but the convergence is obtained when the number of iterations is equal to 2N2 where N is the number of sequences.

4. REFORMALIGN [18]: Using REFORMALIGN, we construct the final alignment indirectly. First, we start by constructing a profile to the initial multiple sequence alignment. Then, we align each sequence to the profile constructed in the first step. Finally, we merge all the sequences alignment in order to obtain the final alignment.

Thus, Refinement algorithms are used in order to enhance a multiple sequence alignment (MSA). Indeed, we start by an initial multiple sequence alignment by using one multiple sequence alignment algorithm. Then, we apply the refinement algorithm to the initial multiple sequence alignment in order to construct a new more accurate multiple sequence alignment having higher score.

2 Block definition

We propose a new algorithm called *Refin-Align* for refining multiple sequence alignment. *Refin-Align* uses a new definition of block. Indeed, a block is defined as a multiple alignment of substrings extracted from a multiple sequence alignment. A block is formed of at least two adjacent columns separated from the initial alignment on both sides by a column formed only of identical characters. Our new definition of blocks is different from the standard definition of blocks, which presents the blocks as substrings delimited by columns containing at least one gap. The blocks are extracted from the initial multiple alignment and then they will be realigned to improve their scores. A block is defined as follow:

- A set of aligned substrings
- having the same size in each sequence
- A block must contain at least two columns
- No substrings formed the block must be formed entirely of gap
- A block must not contains a column having exactly the same character.

3 Refin-Align: New refinement algorithm

The principle of our algorithm is to extract a misaligned blocks from the sequences that distort the multiple alignment and realign them. The *Refin-Align* algorithm allows improving the quality of an initial multiple alignment by iteratively realigning the blocks of the initial multiple alignment. The advantage of our new block definition is to allow

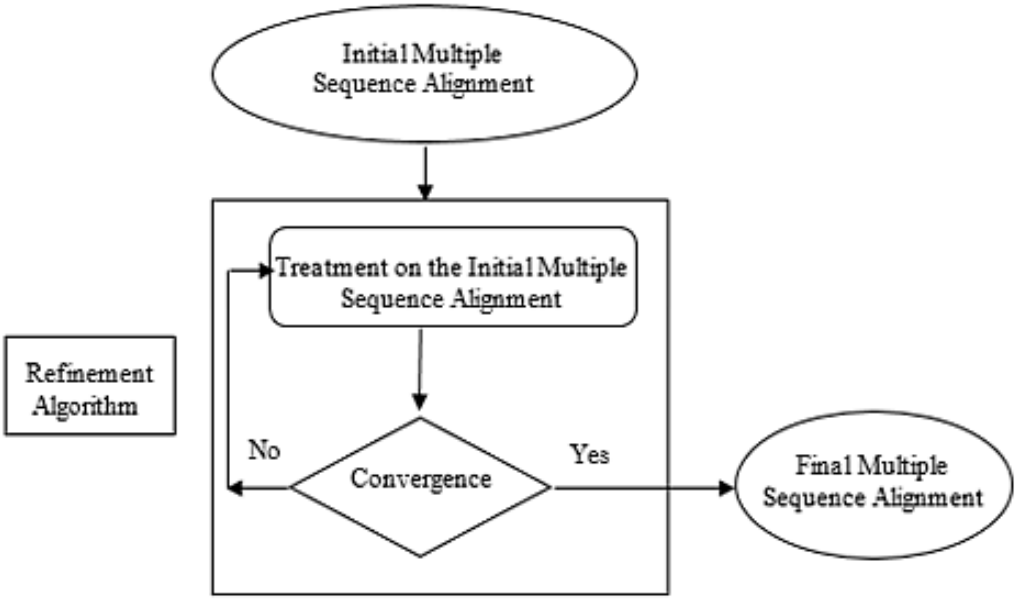


Figure 3: Refinement process.

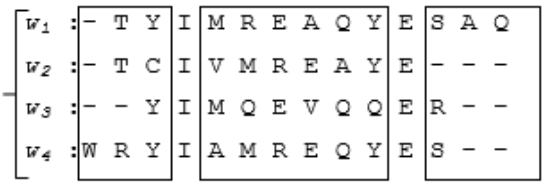


Figure 4: Block extraction.

more possibility for the characters of the initial alignment to be realigned. Our *Refin-Align* algorithm operates as follows:

1. First, we extract the blocks from the initial multiple sequence alignment.

2. Then, we compute the scores of each block. We use the sum of pairs score SP[19]. The SP score correspond to the sum of the scores for all pairs of aligned characters. SP score is computed using this formula:

$$SP(A) = \sum_{i=1}^{L} \sum_{1<k<j<l} s(w_k[i], w_j[i]) \qquad (1)$$

   Where $w_k[i]$ and $w_j[i]$ are the characters in the sequences *k* and *j* that are in the ith column of the alignment *A*,*L* is the length of the alignment *A* and *s* is the score of aligning a pair of characters.

3. Then, we delete gap character from each block and we apply a multiple sequence alignment algorithm Promalign[20] to align these set of new sequences.

4. Finally, we compute the new SP scores. In the case where the scores of the new multiple alignment of blocks are higher than the previous scores, the initial alignment is replaced by the new alignment of blocks obtained.

We repeat this same process, by identifying the new blocks, until we can no longer improve the SP score of each block. The same process is applied for all the blocks of the multiple alignment.

4 Illustrative example

Let be A a multiple sequence alignment of a set of 4 sequences. From this alignment, we extract the blocks B1, B2, B3.

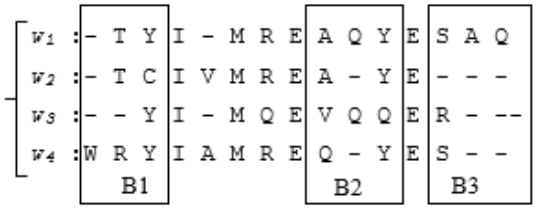


Figure 5: Alignment A contains three blocks B1, B2, B3.

Alignment A contains three blocks B1, B2, B3, we will present the treatment of the second block B2. The same process is repeated for all the blocks.

We compute the SPb, i.e., the SP score of the block B2 before alignment, using the VTML200 Matrix [21]. (* in

```
A Q Y
A - Y
V Q Q
Q - Y
```

Figure 6: Block B2.

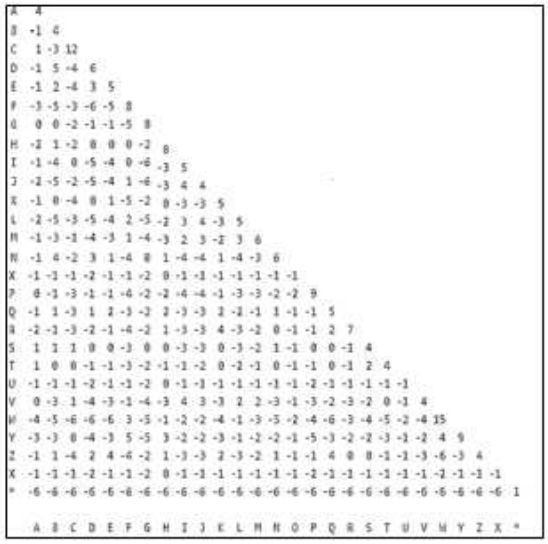


Figure 7: Matrix VTML200.

the Matrix represent the gap '-' character)

SPb:= s(A,A)+2*s(A, V) + 2*s(A, Q) + s(V, Q)+ 4* s(Q, -) + s(Q, Q) + s(-, -) + 3*s(Y,Y)+ 3* s(Y,Q).
SPb:= 0.

Then, we delete gap '-' characters.

```
A Q Y
A Y
V Q Q
Q Y
```

Figure 8: Block B2 without gap.

Then, we realign the block B2, we obtain the following new block.

We compute the SPa score. The SP score of the block B2 after alignment.

SPa:= s(A,A)+2*s(A, V) + 2*s(A, -) + s(V, -) + 3*s(Q, -) +3*s(Q, Q) + 3*s(Y,Y) + 3* s(Y,Q)
SPa:= 1.

The score SPa after alignment is higher than the score SPb before alignment. Thus, we replace the old block B2

```
A Q Y
A - Y
V Q Q
- Q Y
```

Figure 9: Realign the block B2.

by the new block B2 in the initial multiple sequence alignment.

We obtain the following new multiple sequence alignment.

```
w1 :- T Y I - M R E A Q Y E S A Q
w2 :- T C I V M R E A - Y E - - -
w3 :- - Y I - M Q E V Q Q E R - -
w4 :W R Y I A M R E - Q Y E S - -
```

Figure 10: Multiple sequence alignment after refinement.

5 Experimental study

In this section, we present the experimental study realized in order to evaluate the performances of our algorithm. In this experimental study, we use the datasets extracted from several benchmarks. These benchmarks maintain reference multiple sequence alignments constructed in manually or automatically. Moreover these benchmarks continent the scores that allow to compare between the reference multiple sequence alignment in the benchmarks and the test multiple sequence alignment. We used the following scores to compare between the refined multiple sequence alignment obtained using our *Refin-Align* algorithm and the reference multiple sequence alignment in the benchmark.

– (Column Score (CS) [22] is the ratio between the number of correctly aligned columns and the number of all columns whose alignments are known.

$$CS = 1/L * \sum_{i=1}^{L} C_i \quad (2)$$

$C_i = 1$ if all the character of the ith column of the test alignment well aligned in the reference alignment in the benchmark else $C_i = 0$. L the number of column where their alignment are known.

– (Sum of Pairs Score (SPS) [22] is the ratio between the number of correctly aligned pairs of character and the total number of all pairs of character whose alignments are known.

$$SPS = \frac{\sum_{i=1}^{c_t} P_i}{\sum_{r=1}^{c_r} P_r} \quad (3)$$

P_i is the number of pairs of character well aligned in the column i, C_t is the number of column in the alignment test, P_r the total number of all pairs of character whose alignments are known. C_r the number of column in the reference alignment.

We used the Qscore program [5] to compute different scores of the different multiple sequences alignments. Each benchmark uses one notation of the same scores for example BALIBASE uses SPS and CS scores however PREFAB uses respectively Q and TC scores. In our experimental study we used Q and TC scores notations. The datasets used in our experimental study are extracted from the following benchmarks for protein sequences:

1. BALIBASE [23]: This benchmark is the first benchmark dedicated to protein multiple alignment algorithms and contains a number of accurate reference alignments grouped in different references according to the nature of the set of the sequences used. The alignments are constructed based on the superposition of proteins tertiary structures and manual improvement of the results. BALIBASE in the first version contain 5 references, in the last version BALIBASE other references are included. The references are:
   - Reference 1 contains short sequences with different sizes,
   - Reference 2 is composed of sequence families aligned with one, two or three orphan sequences,
   - Reference 3 is composed by groups of sequences having 25% of identity by groups,
   - Reference 4 and 5 are composed by extensions and insertions in the sequences,
   - Reference 6, 7 and 8 are composed by repeat and circular permutation in the sequences.
   - Reference 9 contains motifs in all the sequences.

   BALIBASE uses the CS and SPS.

2. PREFAB [5]: This benchmark is made up of 1932 multiple alignments constructed automatically in the following way: The tertiary structures of two sequences are aligned by using two different superposition methods. A set of 50 homologous sequences is then extracted from databases and a multiple alignment is constructed for the whole set of sequences. PREFAB uses only the Q score that is similar to the SPS score of BALIBASE because the comparison is realized between two aligned sequences, extracted from the reference multiple sequence alignment, and the pairwise alignment of the same sequences extracted from the test multiple sequence alignment

3. OXBENCH [24]: This benchmark is constructed in an automatic way, by aligning known tertiary structures extracted from the Protein Data Bank (PDB) using the AMPS method [25]. OXBENCH uses the Q score and the TC score.

4. HOMSTRAD [26]: It contains 1032 multiple sequences alignments of protein sequences representing different structures and grouped in homologous families.

We assess our program *Refin-Align* using the following methods:

- First, we construct for every dataset an initial set of multiple sequences alignments using the following programs: Clustal Omega, MUSCLE, and MAFFT.
- Then, we compute the column score (CS) [22] and the sum of pairs scores (SPS) [22] before refinement for every multiple sequence alignment in the set.
- Then, we apply our algorithm *Refin-Align* to each multiple sequence alignment in order to obtain the refinement multiple sequence alignment. After that, we compute the column scores (CS) and the sum of pairs scores (SPS) for each multiple sequences alignment after refinement.
- Finally, we compare between the scores obtained before applying our refinement algorithm and those obtained after applying our refinement algorithm.

Scores	Q-scores		T-scores	
	Before	After	Before	After
MAFFT	73,30	**73,42**	52,48	**52,85**
MUSCLE	73,04	**73,82**	54,01	**54,63**
Clustal Omega	70,06	**71,36**	46,70	**47,14**

Table 1: Scores obtained using HOMSTRAD Benchmark

The results of the Program MUSCLE, MAFFT and Clustal Omega are respectively obtained using the program MUSCLE, the online web server of MAFFT and the online web server of Clustal Omega.

These tables below represent the SPS and CS scores obtained.

Table 1 represents the Q-scores and the TC scores obtained before refinement and the scores after refinement on a set of multiple alignment sequence extracted from HOMSTRAD Benchmark.

We benchmarked also our program *Refin-Align* on a set of datasets extracted from OXBENCH Benchmark. Table 2 shows the average of the TC scores and the Q-scores obtained.

We benchmarked also our program *Refin-Align* on several datasets extracted from PREFAB Benchmark.

Scores	Q-scores		T-scores	
	Before	After	Before	After
MAFFT	79,63	**81,63**	70,20	**71,60**
MUSCLE	80,45	**81,74**	70,21	**70,89**
Clustal Omega	84,37	**84,42**	**67,25**	67,04

Table 2: Scores obtained using OXBENCH Benchmark

The comparison of alignments for the PREFAB and the scores computing is different from other benchmarks; this is due to the method of creating this benchmark. Indeed, the reference alignments is a set of pairwise alignment extracted from the multiple sequence alignment. Thus, the Q-scores are computed between the reference pairwise alignment and the test pairwise alignment of the same sequences. In this case, the Q and TC scores are identical.

Table 3 shows the average of Q-scores obtained by applying our refinement *Refin-Align* algorithm on a set of multiple sequence alignment of datasets extracted from PREFAB.

Q-scores	Before	After
MAFFT	**62,07**	62,07
MUSCLE	65,06	**66,04**
Clustal Omega	**64,60**	64,19

Table 3: Scores obtained using PREFAB Benchmark

We also benchmarked our algorithm *Refin-Align* on all the 44 datasets of RV12 reference of BALIBASE and we compute the Q-scores and the TC scores. RV12 represents the reference 1 of the BALIBASE benchmark that contain sequences having between 20% and 40% of identity. Table 4 represents the average scores obtained.

Scores	Q-scores		T-scores	
	Before	After	Before	After
MAFFT	93,71	**93,72**	84,38	**84,40**
MUSCLE	91,55	**91,64**	80,90	**81,06**
Clustal Omega	**90,60**	90,59	79,37	**79,38**

Table 4: Scores obtained using RV12 BALIBASE Benchmark

We note that for several datasets, our refinement algorithm *Refin-Align* can ameliorate the scores of many different multiple sequences alignments obtained by different multiple sequences alignment algorithms. In fact, the refinement multiple sequence alignment after refinement have the best SPS and CS scores for the most datasets used.

6 Conclusion and perspectives

In this paper, we presented a new refinement algorithm for multiple sequence alignment called *Refin-Align*. Our algorithm adopts a new definition of block and use theses blocks to construct a new multiple sequence alignment by realigning these blocks.

We assess our algorithm using different datasets extracted from different benchmarks and using the more efficient multiple sequence alignment algorithms MUSCLE, MAFFT and Clustal Omega. For several datasets, our algorithm can ameliorate the SPS and CS scores for the initial multiple alignment.

In future work, we would like also to compare the results obtained by our program to other refinement programs. We would like also to asses our algorithm on DNA and RNA datasets. We can also improve the scores by using different alignment algorithms to align the blocks in order to obtain the more accurate multiple sequence alignment.

References

[1] Needleman B.S., Wunsch, D. C., (1970) A general method applicable to the search for similarities in the amino-acid sequence of two proteins, *Journal of Molecular Biology*, 48, pp. 443–453. https://doi.org/10.1016/0022-2836(70)90057-4.

[2] Wang, L., Jiang, T., (1994). On the complexity of multiple sequence alignment, *J. Comput. Biol.* 1(4), pp. 337-348. https://doi.org/10.1089/cmb.1994.1.337.

[3] Thompson, J. D., Higgins, D. G., Gibson, T. J., (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleid Acids Research*, 22(22), pp. 4673-4680. https://doi.org/10.1093/nar/22.22.4673.

[4] Notredame, C., Heringa, J., Higgins, D., (2000). T-COFFEE: A novel method for fast and accurate multiple sequence alignments. *J. Molecular Biology*, 302(1), pp. 205-217. https://doi.org/10.1006/jmbi.2000.4042.

[5] Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), pp. 1792-1797. https://doi.org/10.1093/nar/gkh340.

[6] Katoh, K., Kuma, K., Toh, H., Miyata, T., (2013). MAFFT version 7: Improvement in accuracy of

multiple sequence alignment. *Molecular Biology and Evolution*, 30(4), pp. 772-780. https://doi.org/10.1093/molbev/mst010.

[7] Yongtao, Y., Cheung, David, W., Wang, Yadong W., Yin, S. M., Zhang, Q., Lam, T. W., Ting, H. F.,.(2013). GLProbs: Aligning Multiple Sequences Adaptively. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(1), pp. 67-78. https://doi.org/10.1109/TCBB.2014.2316820.

[8] Sievers, Fabian, Higgins, Desmond, G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences, *Multiple Sequence Alignment Methods*, (1079), pp, 105-116. https://doi.org/10.1007/978-1-62703-646-7_6.

[9] Mokaddem A. Elloumi, M. (2013). New distances for improving progressive alignment algorithm. *Advances in Computing and Information Technology*, (177), pp. 243-251. https://doi.org/10.1007/978-3-642-31552-7_26.

[10] Kimura, M. (1983). The Neutral Theory of Molecular Evolution. *Cambridge University Press*, Cambridge. https://doi.org/10.1017/CBO9780511623486.

[11] Sneath, P., Sokal, R., (1973). Numerical Taxonomy. *San Francisco Freeman*, pp. 230-234. https://doi.org/10.2307/2529664.

[12] Saitou, N., Nei, M..(1987). The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol*, 4(4), pp. 406-425. https://doi.org/10.1093/oxfordjournals.molbev.a040454.

[13] Naznin, F, Sarker. R, Essam D, (2012). Progressive Alignment Method Using Genetic Algorithm for Multiple Sequence Alignment. *IEEE Transactions on Evolutionary Computation*, 16(5), pp. 615-631. https://doi.org/10.1109/TEVC.2011.2162849.

[14] Behera, N., Jeevitesh, M. S., Josea, J., Kant, K., Dey, A., Mazher, J., (2017). Higher accuracy protein multiple sequence alignments by genetic algorithm. *Procedia Computer Science*, 108, pp. 1135-1144. https://doi.org/10.1016/j.procs.2017.05.100.

[15] J. D. Thompson, J. C. Thierry, O. Poch (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*, 19(9), pp 1155-1161. https://doi.org/10.1093/bioinformatics/btg133.

[16] Chakrabarti, S., Lanczycki, C. J., Panchenko, A. R., Przytycka, T. M., Thiessen, P. A., Bryant, S. H., (2006). Refining multiple sequence alignments with conserved core regions, *Nucleic Acids Res*,34(9), pp. 2598-2606. https://doi.org/10.1093/nar/gkl274.

[17] Wallace, I.M., Blackshields, G., Higgins, D.G., (2005). Multiple sequence alignments. *Current Opinion in Structural Biology*, 15, pp. 261-266. https://doi.org/10.1016/j.sbi.2005.04.002.

[18] Lyras, Dimitrios P., Metzler, Dirk. (2014) ReformAlign: improved multiple sequence alignments using a profile-based meta-alignment approach. *BMC bioinformatics*, 15(1), pp. 265. https://doi.org/10.1186/1471-2105-15-265.

[19] Altschul, S.F.(1989). Gap costs for multiple sequence alignment. *J Theor Biol.*, 138(3), pp. 297–309. https://doi.org/10.1016/S0022-5193(89)80196-1.

[20] Mokaddem, A.,Hadj, A. B., Elloumi, M., (2018). Pro-malign: Multiple Sequence Alignment Algorithm using Approached Profile. *Journal of software*, 13 (1), pp. 57-65. https://doi.org/10.17706/jsw.13.1.57-65.

[21] Muller, T., Spang. R, Vingron. M, (2002). Estimating amino acid substitution models: a comparison of dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol Biol Evol*, 19(1), pp. 8-13. https://doi.org/10.1093/oxfordjournals.molbev.a003985.

[22] Thompson, J.D., Plewniak, F., Poch, O. (1999). BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15(1): pp. 87-88. https://doi.org/10.1093/bioinformatics/15.1.87.

[23] Thompson, J.D.,, Koehl P., Ripp R., Poch O.(2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *PROTEINS: Structure, Function, and Bioinformatics*, 61(1), pp. 127-136. https://doi.org/10.1002/prot.20527.

[24] Raghava, G. P., Searle, S. M., Audley, P. C., Barbe,r J. D., Barton, G. J., (2003). OXBENCH: a benchmark for evaluation of protein multiple sequence alignment

accuracy. *BMC Bioinformatics*, 4(1), pp. 47. https://doi.org/10.1186/1471-2105-4-47.

[25] Barton, G.J., Sternberg, M.J.: (1987). A strategy for the rapid multiple alignment of protein sequences, confidence levels from tertiary structure comparisons, *J Mol Biol*, 198(2), pp. 327-337. https://doi.org/10.1016/0022-2836(87)90316-0.

[26] Stebbings, L. A. and Mizuguchi, K., (2004). HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Research*, 32, pp. 203–207. https://doi.org/10.1093/nar/gkh027.

A Novel Approach to Fuzzy-Based Facial Feature Extraction and Face Recognition

Aniruddha Dey
Department of Computer Science & Engineering, Jadavpur University, Kolkata, India.
E-mail: anidey007@gmail.com

Manas Ghosh
Department of Computer Application, RCCIIT, Kolkata, India
E-mail: manas.ghosh@rcciit.org

Keywords: fuzzy G-2DLDA, fuzzy set theory, Fk-NN, RBFNN based classifier, membership degree matrix

Received: December 29, 2017

Abstract: Generalized two-dimensional Fisher's linear discriminant (G-2DFLD) is an effective feature extraction technique that maximizes class separability along row and column directions simultaneously. In this paper, we have presented a fuzzy-based feature extraction technique, named fuzzy generalized two-dimensional Fisher's linear discriminant analysis (FG-2DLDA) method. The FG-2DLDA is extended version of the G-2DFLD method. In this study, we also have demonstrated the face recognition using the presented method with radial basis function (RBF) as a classifier. In this context, it is to be noted that the fuzzy membership matrix for the training samples is computed by means of fuzzy k-nearest neighbour (Fk-NN) algorithm. The global mean and class-wise mean training images are generated by combining the fuzzy membership values with the training samples. These mean images are used to compute the fuzzy intra- and inter-class scatter matrices along x- and y-directions. Finally, by solving the Eigen value problems of these scatter matrices, we find the optimal fuzzy projection vectors, which actually used to generate more discriminant features. The presented method has been validated over three public face databases using RBF neural network and establish that the proposed FG-2DLDA method provides favourable recognition rates than some contemporary face recognition methods.

Povzetek: V prispevku je opisana metoda dvodimenzionalne Fisherjeve linearne diskriminacijske analize na osnovi mehkih množic (FG-2DLDA).

1 Introduction

Facial feature extraction technique has developed as a popular research area in last 20 years in the field of computer vision, and machine learning [1- 6]. Very popular linear methods include principal component analysis (PCA) [5- 6], linear discriminant analysis (LDA) [7] and their variants, which use Eigen faces and/or Fisherfaces to compute features, fall under this category. In particular, PCA maximizes the total scatter matrix across all face images. However, undesirable variations caused by lighting, facial expression and other factors are retained through PCA techniques [6]. Many researchers argue that the PCA techniques do not provide any information for class discrimination; only perform dimension reduction [6, 7]. The LDA has been proposed as a better alternative to the PCA to provide class discrimination information [8, 9]. The main objective of the LDA is to find best discrimination of vectors among the classes by maximizing the between-class differences and minimizing the within-class ones [8]. The disadvantage of LDA technique is that, it suffers from the "small sample size (*SSS*)" problem [9]. The aforementioned problem mainly occurs in case of few numbers of sample than the sample dimension. The dimension of face images is generally very high; as a results, the within-class scatter matrix become singular that makes the FLD method infeasible. The *SSS* problem in LDA can be solved by sampling down the face images into smaller size [10]. LDA is one of the most important linear approaches for feature extraction which maximizes the ratio of the between-and within-class scatter matrix. However, for a task with very high-dimensional facial images, LDA method may suffer from the problem of singularity. To solve this problem, PCA has been applied to reduce the dimensions of the high dimensional vector space before employing the LDA method [11]. While the PCA seeks projections which are optimal for image reform from a low dimensional space, it may remove dimensions that contain discriminant information required for face recognition. R-LDA method introduces to solve the singularity problem [12]. The main drawback of R-LDA is that the dimensionality of covariance matrix is often more than ten thousand. It is not useful for R-LDA to procedure such large covariance matrix, when the computing platform is not sufficiently powerful. Huang et al. [13] introduced a more efficient null space IDA method. The key idea of this technique is that the with-class scatter matrix (S_w) is more effective for calculating discriminant feature, whereas, between-class scatter matrix (S_b) is useless. Though, the method is

often criticized for the high storage requirement and computational cost in facial feature extraction and recognition. Chen et al. [8] claimed that eigenvectors corresponding to eigenvalues equal to zero of S_w contain the maximum discriminant information. Yu and Yang [14] proposed a direct linear discriminant analysis method by diagonal the between- and within-class scatter matrix. It is well-known that between- and within-class scatter are two important measures of the separability of the projected samples. Independent component analysis (ICA) is also proposed as an effective feature extraction technique [15]. ICA computes discriminant features from covariance matrix by considering high-order statistics. The two-dimensional PCA (2DPCA) directly works on the 2D image matrices and found to be computationally efficient and more superior for face recognition and reconstruction than PCA [16]. Two-dimensional FLD (2DFLD) method maximizes the class separability in one direction (row or column) at a time [17]. The significant characteristic of 2DFLD method is that it directly works on the 2D image matrices. The projection vectors are extracted from the row and, by the G-2DFLD method [18]. The discriminant feature matrices are found by linearly projecting an image matrix on aforementioned directions. Therefore, the discriminative information is maximized by this method among the classes while minimizing it column direction of the training images simultaneously within a class [18]. To increase its pertinence, many LDA extensions, such as direct LDA [19], complete LDA [20], LDA/QR [21] or LDA/GSVD [22], have been developed in the last decades. These extensions try to preserve the same validation and overcome singularity problems either by first projecting the problem in a convenient subspace, using alternative indirect or approximate optimizations.

Very recently, several researchers presented fuzzy-based methods, such as fuzzy *k*-nearest neighbour (F*k*-NN) [23], fuzzy two dimensional Fisher's linear discrimination (F-2DFLD) [25], fuzzy maximum scatter difference (F-MSD) [28], fuzzy two dimensional principal component analysis (F-2DPCA) [32], fuzzy two dimensional linear discriminant analysis (GPG-2DLDA), Generalized multiple maximum scatter difference (GMMSD) [33], fuzzy local mean discriminant analysis (FLMDA) [36], and fuzzy linear regression discriminant projection (FLRDP) [37] for feature extraction. Keller et al. (1985) presented the fuzzy *k*-nearest neighbour (F*k*-NN) approach, which fuzzifies the class assignment [23]. This method, popularly known as fuzzy Fisherface [24] (Fuzzy-FLD), which incorporates the fuzzy membership grades into the within- and between-class scatter matrices for binary labelled patterns to extract features and are used for face recognition [25]. The fuzzy 2DFLD (F-2DFLD) is an extension of the fuzzy Fisherface [26]. The scatter matrices were redefined by introducing membership values into each training sample. Yang *et al.* proposed feature extraction using fuzzy inverse FDA [26]. The F*k*-NN was also incorporated in fuzzy inverse FDA for calculating membership degree matrices. The F*k*-NN is used to calculate the membership matrix, which is incorporated within the definition of between class and within class scatter matrix [26]. Reformative LDA method is used along with the F*k*-NN method to redefine the scatter matrices [27]. A weighted maximum scatter difference algorithm is used for face recognition [28]. Fuzzy LDA algorithm is derived by incorporating the fuzzy membership into learning and random walk method is introduced to reduce the effect of outliers [29]. Fuzzy set theory is integrated with the scatter difference discriminant criterion (SDDC) algorithm where F*k*-NN method is used to compute the membership grade which is utilized to redefine the scatter matrices [30]. Fuzzy maximum scatter difference model is proposed where F*k*-NN is used to calculate the membership degree matrix of training sample [31]. The Fuzzy 2DPCA method was introduced where F*k*-NN method is applied to compute the membership matrix for training sample which was utilized to obtain fuzzy mean of each class. The average of the mean was calculated to define the scatter matrices [32]. Generalized multiple maximum scatter difference discriminant criterion has been introduced for effective feature extraction and classification [33]. Gaussian probability distribution information was incorporated in defining of between class and within class scatter matrices [34]. The membership grade and label information were used to define the scatter matrices [35]. Fuzzy local mean discriminant analysis was employed to construct the scatter matrices by redefining the fuzzy local class means [36]. Fuzzy linear regression discriminant projection method is proposed to compute the fuzzy membership grade for each sample and incorporated in the definition of within class and between class scatter matrices [37].

In the proposed method, we have incorporated the fuzzy membership values in different classes which are computed from the training images (samples). To obtain the membership degrees of each training sample, we have used the fuzzy *k*-NN and used them for calculating the global and class-wise mean training image matrices. Finally, fuzzy scatter matrices (between and within) are computed distinctly in row and column wise direction. To solve the eigenvalue problem of aforementioned scatter matrices, the features are extracted.

The remaining sections of this paper are organized as follows. In Section 2, we give brief overview of G-2DFLD method. In Section 3, we propose a novel method for feature extraction based on G-2DFLD method, called FG-2DLDA method. The simulation results on three public face image datasets are demonstrated in Section 4. Concluding remarks is given Section 5.

2 Brief summary of the generalized 2DFLD method

Our presented technique is extended version of the G-2DFLD feature extraction technique [18]. G-2DFLD method is briefly presented in this section.

Let, the face images are of $r \times s$ dimension which are represented in the form of 2D vectors $\boldsymbol{X}_i$ $(i = 1, 2, \dots, N)$. The total number of "C" classes comprises N

face images. The c^{th}class is represented by C_c having total samples of N_c and also satisfying the condition $(\sum_{c=1}^{C} N_c = N)$. Given an image $\boldsymbol{X}$, the G-2DFLD-based 2D feature matrix $\boldsymbol{Y}$ is generated by the following linear transformation:

$$\boldsymbol{Y} = (\boldsymbol{P}_{opt})^T \boldsymbol{X} (\boldsymbol{Q}_{opt}) \tag{1}$$

where $\boldsymbol{P}_{opt}$ and $\boldsymbol{Q}_{opt}$ are the two optimal projection matrices.

The two Fisher's criteria (objective function) along row and column direction ($J(\boldsymbol{P})$, $J(\boldsymbol{Q})$) have been expressed as stated below:

$$\begin{cases} J(\boldsymbol{P}) = \boldsymbol{P}^T \boldsymbol{G}_{rb} \boldsymbol{P} \times (\boldsymbol{P}^T \boldsymbol{G}_{rw} \boldsymbol{P})^{-1} \\ and \\ J(\boldsymbol{Q}) = \boldsymbol{Q}^T \boldsymbol{G}_{cb} \boldsymbol{Q} \times (\boldsymbol{Q}^T \boldsymbol{G}_{cw} \boldsymbol{Q})^{-1} \end{cases} \tag{2}$$

The optimal projection vectors $\boldsymbol{P}_{opt}$ and $\boldsymbol{Q}_{opt}$ can be obtained by finding the normalized eigenvalues the eigenvectors of $\boldsymbol{G}_{rb}\boldsymbol{G}_{rw}^{-1}$ and $\boldsymbol{G}_{cb}\boldsymbol{G}_{cw}^{-1}$, respectively. The eigenvalues are sorted in descending order and the eigenvectors are also rearranged accordingly [18]. The optimal projection (eigenvector) matrix $\boldsymbol{P}_{opt}$ and $\boldsymbol{Q}_{opt}$ can be stated as follows:

$$\begin{cases} \boldsymbol{P}_{opt} = \arg\max_{P} |\boldsymbol{G}_{rb}\boldsymbol{G}_{rw}^{-1}| \\ \quad = [p_1, p_2, \dots, p_u] \\ and \\ \boldsymbol{Q}_{opt} = \arg\max_{Q} |\boldsymbol{G}_{cb}\boldsymbol{G}_{cw}^{-1}| \\ \quad = [q_1, q_2, \dots, q_v] \end{cases} \tag{3}$$

The between-class and within-class scatter matrices along row direction ($\boldsymbol{G}_{rb}$ and $\boldsymbol{G}_{rw}$) and column direction ($\boldsymbol{G}_{cb}$ and $\boldsymbol{G}_{cw}$) are computed as follows :

$$\begin{cases} \boldsymbol{G}_{rb} = \sum_{c}^{C} N_c (\boldsymbol{m}_c - \boldsymbol{m})(\boldsymbol{m}_c - \boldsymbol{m})^T \\ and \\ \boldsymbol{G}_{rw} = \sum_{c}^{C} \sum_{i \in c}^{N} (\boldsymbol{X}_i - \boldsymbol{m}_c)(\boldsymbol{X}_i - \boldsymbol{m}_c)^T \end{cases} \tag{4a}$$

$$\begin{cases} \boldsymbol{G}_{cb} = \sum_{c}^{C} N_c (\boldsymbol{m}_c - \boldsymbol{m})^T (\boldsymbol{m}_c - \boldsymbol{m}) \\ and \\ \boldsymbol{G}_{cw} = \sum_{c}^{C} \sum_{i \in c}^{N} (\boldsymbol{X}_i - \boldsymbol{m}_c)^T (\boldsymbol{X}_i - \boldsymbol{m}_c) \end{cases} \tag{4b}$$

In above expression, the global mean training image $\left(\boldsymbol{m} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i\right)$ and class-wise mean training image $\left(\boldsymbol{m}_c = \frac{1}{N_c}\sum_{i=1}^{N} \boldsymbol{X}_i \,|\boldsymbol{X}_i \in C_c\right)$ are calculated. The dimensions of the row-wise scatter matrices ($\boldsymbol{G}_{rb}$ *and* $\boldsymbol{G}_{rw}$) and the column-wise scatter matrices ($\boldsymbol{G}_{cb}$ *and* $\boldsymbol{G}_{cw}$) are found to be $r \times r$ and $s \times s$, respectively.

3 Proposed fuzzy generalized two-dimensional linear discriminant analysis (FG-2DLDA) method

Human faces are highly susceptible to vary under different environmental conditions, such as illumination, pose, etc. As a result, sometimes, images of a person may look alike to that of a different person. In addition, variability among the images of a person may differ quite significantly. The proposed FG-2DLDA method is basically based on the concept of fuzzy class assignment, where a face image belongs to different classes as characterized by its fuzzy membership values. The idea of fuzzification using fuzzy *k*-nearest neighbour (F*k*-NN) was conceived by Keller et al. and found to be more effective [23]. In the present study, we have used the F*k*-NN for generating fuzzy membership values for training images; resulting a fuzzy membership matrix. The fuzzy membership values are incorporated with the training images to obtain global and class-wise mean images, which in turn used to form fuzzy (between- and within-class) scatter matrices. Therefore, these scatter matrices yield useful information regarding association of each training image into several classes. The optimal fuzzy 2D projection vectors are obtained by solving the eigenvalue problems of these scatter matrices. Finally, the FG-2DLDA-based features are extracted by projecting a face image onto these optimal fuzzy 2D projection vectors. The different steps of the FG-2DLDA method are presented in details in the following sub-sections.

3.1 Generation of membership matrix by fuzzy k-nearest neighbour (Fk-NN)

Let, there are *C* classes and *N* training images; each one is represented in the form of 2D vectors $\boldsymbol{X}_i$ $(i = 1, 2, \dots, N)$. A fuzzy *k*-NN-based decision algorithm has been performed for assigning membership values (degree) to the training images [23, 24]. This Fuzzy k-Nearest Neighbour *(Fk-NN)* method redefines the membership values of the labelled face images. When, all of the neighbours belong to the i^{th} class which is equal to the class of j^{th} image under consideration, then $n_{ij} = k$ and μ_{ij} returns 1, making membership values for the other classes as zero. In addition, μ_{ij} also satisfies two obvious properties $\left(\sum_{i=1}^{C} \mu_{ij} = 1 \; and \; 0 < \sum_{j=1}^{N} \mu_{ij} < N\right)$. So, the fuzzy membership matrix *U* using the F*k*-NN can be demonstrated as given below:

$$U = [\mu_{ci}]; c = 1, 2, 3, \dots, C; \quad i = 1, 2, 3, \dots, N \tag{5}$$

3.2 Fuzzy generalized two dimensional linear discriminant analysis (FG-2DLDA) algorithm

FG-2DLDA methods has employed the fuzzy membership values with the training images and redefine the scatter matrices along row and column directions. Finally, the optimal fuzzy projection vectors are generated by solving the eigenvalue problems of these

scatter matrices. Let the training set contains N images of C classes (subjects) and each one is denoted as $\boldsymbol{X}_i$ ($i = 1, 2, 3 \ldots, N$) having dimension as $r \times s$. The c^{th} class C_c, has total N_c images and satisfies $\sum_{c=1}^{C} N_c = N$.

For an image $\boldsymbol{X}$, the FG-2DLDA-based features in the form of 2D matrix of size $u \times v$ is generated by projecting it onto the optimal fuzzy projection matrices and can be achieved by the following linear transformation as defined below:

$$\boldsymbol{Y}^f = \left(\boldsymbol{P}_{opt}^f\right)^T \boldsymbol{X} \left(\boldsymbol{Q}_{opt}^f\right) \tag{6}$$

The Fisher's criteria (objective function) $J^f(\boldsymbol{P})$ and $J^f(\boldsymbol{Q})$ along row and column directions are defined as follows:

$$\begin{cases} J^f(\boldsymbol{P}) = (\boldsymbol{P}^f)^T \boldsymbol{G}_{rb}^f \boldsymbol{P}^f \times \left\{(\boldsymbol{P}^f)^T \boldsymbol{G}_{rw}^f \boldsymbol{P}^f\right\}^{-1} \\ \quad and \\ J^f(\boldsymbol{Q}) = (\boldsymbol{Q}^f)^T \boldsymbol{G}_{cb}^f \boldsymbol{Q}^f \times \left\{(\boldsymbol{Q}^f)^T \boldsymbol{G}_{cw}^f \boldsymbol{Q}^f\right\}^{-1} \end{cases} \tag{7}$$

The ratio is maximized in the above equations (7) when the column vectors of the projection matrix $\boldsymbol{P}^f$ and $\boldsymbol{Q}^f$ are the eigenvectors of $\boldsymbol{G}_{rb}^f \left(\boldsymbol{G}_{rw}^f\right)^{-1}$ and $\boldsymbol{G}_{cb}^f \left(\boldsymbol{G}_{cw}^f\right)^{-1}$, respectively. The fuzzy optimal projection matrix $\boldsymbol{P}_{opt}^f$ and $\boldsymbol{Q}_{opt}^f$ are obtained by finding the eigenvectors of $\boldsymbol{G}_{rb}^f \left(\boldsymbol{G}_{rw}^f\right)^{-1}$ and $\boldsymbol{G}_{cb}^f \left(\boldsymbol{G}_{cw}^f\right)^{-1}$ corresponding to the u and v largest eigenvalues, respectively. The fuzzy optimal projection matrices $\boldsymbol{P}_{opt}^f$ and $\boldsymbol{Q}_{opt}^f$ can be represented as follows:

$$\begin{cases} \boldsymbol{P}_{opt}^f = \arg\max_{Pf} \left| \boldsymbol{G}_{rb}^f \left(\boldsymbol{G}_{rw}^f\right)^{-1} \right| \\ \qquad = [p_1, p_2, \ldots, p_u] \\ and \\ \boldsymbol{Q}_{opt}^f = \arg\max_{Qf} \left| \boldsymbol{G}_{cb}^f \left(\boldsymbol{G}_{cw}^f\right)^{-1} \right| \\ \qquad = [q_1, q_2, \ldots, q_v] \end{cases} \tag{8}$$

where $\{p_i | i = 1, 2, \ldots, u\}$ is the set of normalized eigenvectors of $\boldsymbol{G}_{rb}^f \left(\boldsymbol{G}_{rw}^f\right)^{-1}$ corresponding to u largest eigenvalues $\{\lambda_i | i = 1, 2, \ldots, u\}$ and $\{q_j | j = 1, 2, \ldots, v\}$ is the set of normalized eigenvectors of $\boldsymbol{G}_{cb}^f \left(\boldsymbol{G}_{cw}^f\right)^{-1}$ corresponding to v largest eigenvalues $\{\alpha_j | j = 1, 2, \ldots, v\}$.

The four (within- and between- class) fuzzy scatter matrices ($\boldsymbol{G}_{rb}^f$, $\boldsymbol{G}_{rw}^f$, $\boldsymbol{G}_{cb}^f$, $\boldsymbol{G}_{cw}^f$) along the row and column directions are defined as follows:

$$\begin{cases} \boldsymbol{G}_{rb}^f = \frac{1}{N} \sum_{c}^{C} N_c^f (\bar{\boldsymbol{m}}_c - \bar{\boldsymbol{m}})(\bar{\boldsymbol{m}}_c - \bar{\boldsymbol{m}})^T \\ and \\ \boldsymbol{G}_{rw}^f = \frac{1}{N} \sum_{c}^{C} \sum_{i \in c}^{N} (\boldsymbol{X}_i - \bar{\boldsymbol{m}}_c)(\boldsymbol{X}_i - \bar{\boldsymbol{m}}_c)^T \end{cases} \tag{9a}$$

$$\begin{cases} \boldsymbol{G}_{cb}^f = \frac{1}{N} \sum_{c}^{C} N_c^f (\bar{\boldsymbol{m}}_c - \bar{\boldsymbol{m}})^T (\bar{\boldsymbol{m}}_c - \bar{\boldsymbol{m}}) \\ and \\ \boldsymbol{G}_{cw}^f = \frac{1}{N} \sum_{c}^{C} \sum_{i \in c}^{N} (\boldsymbol{X}_i - \bar{\boldsymbol{m}}_c)^T (\boldsymbol{X}_i - \bar{\boldsymbol{m}}_c) \end{cases} \tag{9b}$$

Where fuzzy membership degrees are integrated into the training images to get fuzzy global mean image $\left(\bar{\boldsymbol{m}} = \frac{\sum_{c=1}^{C} \sum_{i=1}^{N} \boldsymbol{\mu}_{ci} \boldsymbol{X}_i}{\sum_{c=1}^{C} \sum_{i=1}^{N} \boldsymbol{\mu}_{ci}}\right)$ and fuzzy class-wise mean images $\left(\bar{\boldsymbol{m}}_c = \frac{\sum_{i=1}^{N} \boldsymbol{\mu}_{ci} \boldsymbol{X}_i}{\sum_{i=1}^{N} \boldsymbol{\mu}_{ci}};\ c = 1, 2, 3, \ldots, C\right)$. It may be also noted that the size of the $\left(\boldsymbol{G}_{rb}^f$ and $\boldsymbol{G}_{rw}^f\right)$ scatter matrices is $r \times r$; whereas for the $\boldsymbol{G}_{cb}^f$ and $\boldsymbol{G}_{cw}^f$ scatter matrices it is $s \times s$.

4 Simulation results and discussion

We have assessed the performance of the proposed FG-2DLDA on three publicly available databases namely, FERET [39, 40] AT&T [41], and UMIST [42]. The equation for calculating the recognition rate is represented below:

$$R_{avg} = \frac{\sum_{i=1}^{q} n_{cls}^i}{q \times n_{tot}} \tag{10}$$

where, q denotes total number of experimental runs. Correct recognition number in the i^{th} run is represented by n_{cls}^i. n_{tot} indicates the whole number of test face images.

FERET face database is used to evaluate the FG-2DLDA method under several facial expressions, pose and lighting conditions. AT&T and UMIST database are used to access the presented method under the condition of minor variations of rotation and scaling. In these experiments, we have used a RBFNN classifier due to its superiority and simplicity over the other types of neural networks. As discussed in Section 3 of the proposed FG-2DLDA method, the experiments are performed to validate our claim. The FG-2DLDA algorithm is implemented in C programming language on the Linux operational system with Intel Core i5 (2.4 GHz) and DDR3 (8 GB, 1333 MHz). The suggested method is evaluated on a subset of ***FERET*** face database [39, 40]. The database consists of 1400 images of 200 individuals and each individual is having 7 images. The images differ in facial expression, illumination and pose. In our study, the facial portion of each original image was lopped and resized to 80×80 pixels based on the location of the eyes. Here, the values of s are taken as 2, 3 and 4

and our method is tried out 10 times with each value of s with the different training sets and test sets. Some examples of images of a person are shown in Fig. 1 (i). A set of 400 images of 40 persons comprise ***AT&T face database.*** There are 10 dissimilar images for each person. In our present study, from the set of images for each person, *s* images are picked out in random from the database to generate the training set and remaining (10-*s*) images are considered as the test set. Hence, a distinct set of images encompasses the training and test set. 3, 4, 5, and 6 are taken as the values of *s* to form different pairs of training and test sets. Some examples of images of a individual are shown in Fig. 1 (ii). A total of 575 grey-scaled images of 20 different individuals covering a variety of race, sex, and appearance is contained in the multi-view ***UMIST*** database. The Face database of images per individual varies from 19 to 48 images. In recent studies, we have diminished each image to 112 × 92 pixels. Fig. 1 (iii) shows one person face image from the database.

Figure 1: Some pictures of a person from the (i) FERET, (ii) AT&T, and (iii) UMIST face databases.

The experiments are repeated 10 times for each value of *s* with different training set and test sets on the FERET face database. Here, we choose *s* = 2, 3, 4 images from each subject at random for training and remaining (7-*s*) images are employed for testing. The proposed method is evaluated with feature matrices sizes from 6×6

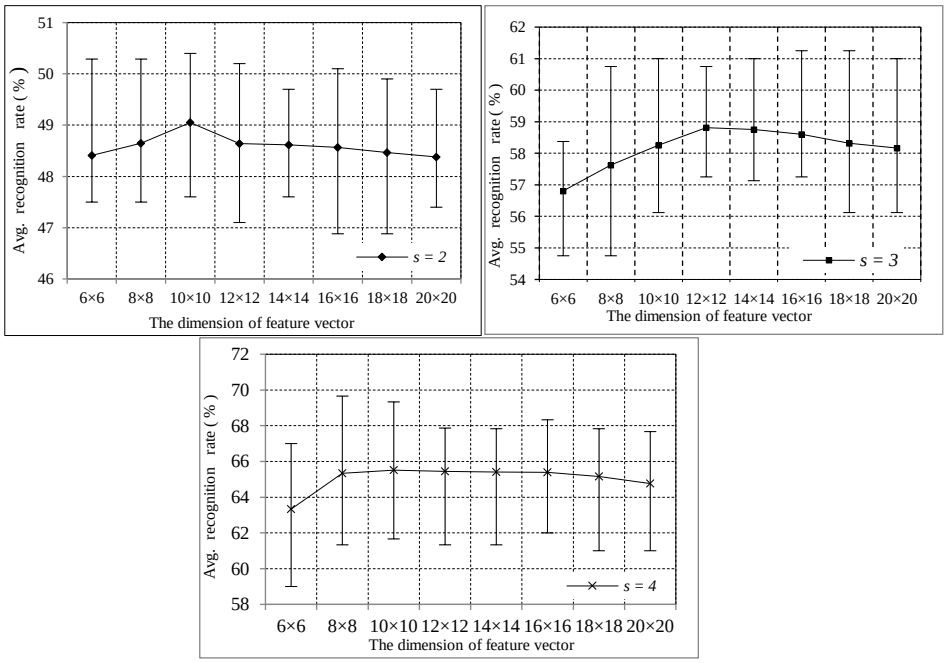


Figure 2: Minimum, maximum and average recognition rates of the FG-2DLDA method for different values of *s* by varying feature size on the FERET face database.

Table 2: Comparison in terms of average recognition rates (%) obtained from different methods on the FERET face database.

Method	Average recognition rates		
	s = 2	*s* = 3	*s* = 4
FG-2DLDA	**49.05** (10×10)	**58.81** (12×12)	**65.51** (10×10)
F-2DFLD [22]	48.88 (40×8)	-	-
MMSD ($\theta = 0.4$) [28]	-	52.6 -	55.81 -
MSD ($\theta = 0.4$) [28]	-	50.5 -	53.68 -
FMSD ($\theta = 0.4$) [28]	-	53.46 -	56.9 -
Alternative-2DPCA [35]	48.31 (112×20)	53.21 (112×20)	53.97 (112×20)
(2D) 2PCA [35]	47.70 (112×20)	52.36 (112×20)	55.45 (112×20)
2DPCA [35]	47.12 (112×20)	52.66 (112×20)	55.20 (112×20)

*Highest recognition rates are indicated by the bold values.

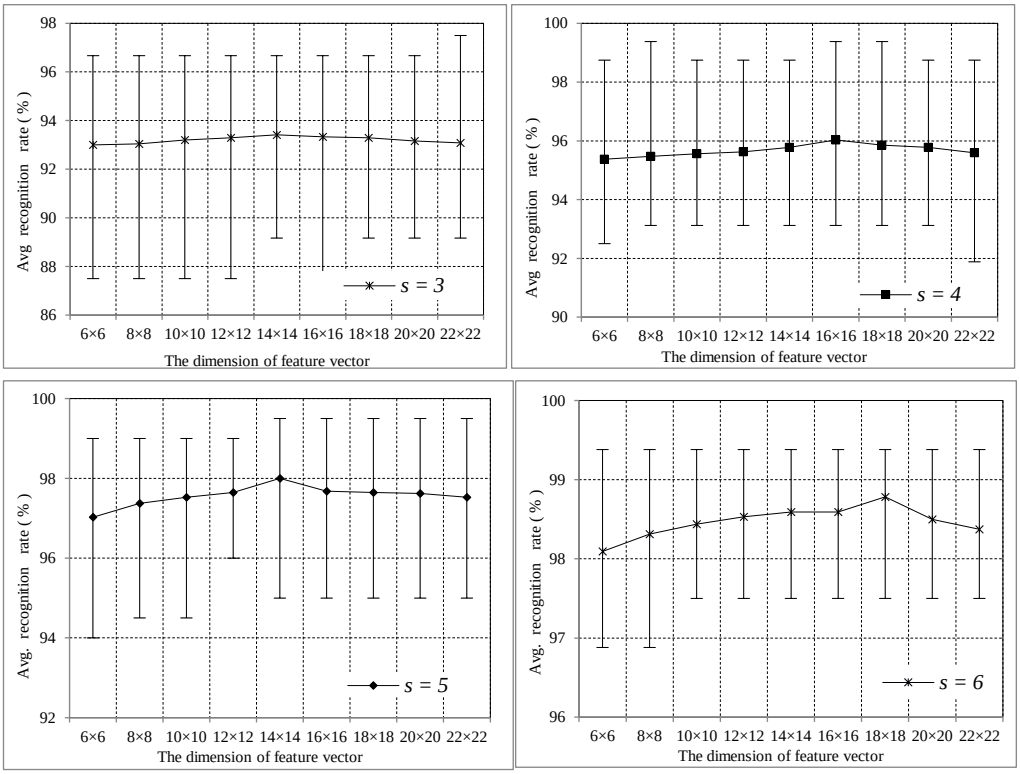


Figure 3: Minimum, maximum and average recognition rates of the FG-2DLDA method for different values of *s* by varying feature size on the AT&T face database.

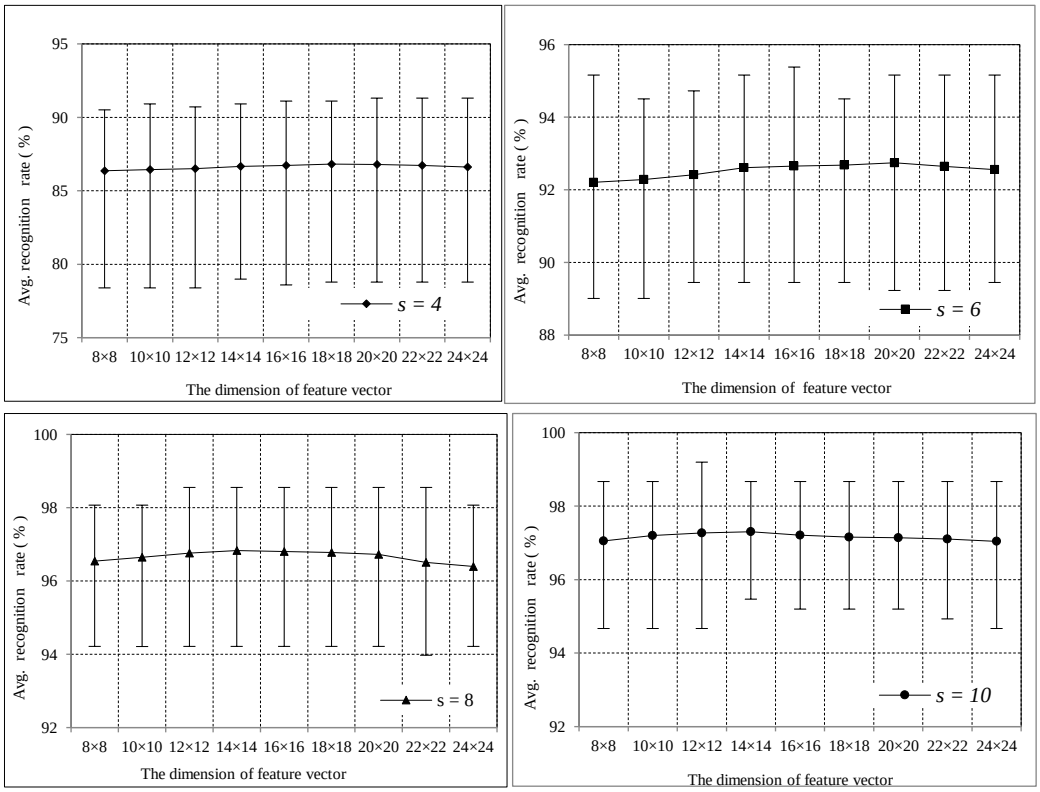


Figure 4: Minimum, maximum and average recognition rates of the FG-2DLDA method for different values of *s* by varying feature size on the UMIST face database.

to 20×20 using RBFNN as a classifier. Fig. 2 demonstrate the minimum, maximum and average recognition rates (%) of the FG-2DLDA method for different values of (*s*= 2, 3 and 4) by varying feature size.

We have compared the performance of the proposed FG-2DLDA method with other competent related methods. FG-2DLDA method extracts discriminative feature by calculating the within class and between class

Table 4: Comparison in terms of average recognition rates (%) obtained from different methods on the UMIST face database.

Method	Average recognition rates			
	s = 4	s = 6	s = 8	s = 10
FG-2DLDA	**86.81** (18×18)	**92.75** (20×20)	**96.83** (14×14)	**97.30** (14×14)
G-2DFLD [18]	86.22 (14×14)	92.28 (14×14)	95.54 (14×14)	96.92 (18×18)
F-LDA [29]	84.5 (19)	-	-	92.01 (19)
MF-LDA [29]	85.38 (19)	-	-	92.53 (19)
2DFLD [18]	86.12 (112×14)	92.16 (112×14)	95.25 (112×14)	96.55 (112×18)
2DPCA [18]	85.70 (112×14)	91.91 (112×14)	95.07 (112×14)	96.60 (112×18)
RF-LDA [29]	84.8 (19)	-	-	92.38 (19)
PCA [18]	80.72 (60)	86.53 (60)	94.01 (60)	95.11 (60)

scatter matrix in row and column direction. Thus, the results again demonstrate the superiority of the FG-2DLDA method over other methods.

In this study, we have validated the performance of our method with 20 different pairs of training and test sets for each value of s on the AT&T face database. Since the present method considers that a face image may simultaneously belong to different classes with possibly different membership values, the class-wise mean images may differ from the actual ones. Fig. 3 Minimum, maximum and average recognition rates of the FG-2DLDA method for different values of s by varying feature size on the AT&T face database. The proposed method yields the best average recognition rates of 93.41% (14×14), 96.08% (16×16), 98.08% (14×14), and 98.68% (18×18) for s = 3, 4, 5, and 6, respectively.

Table 3 demonstrates the best average recognition rates achieved by this algorithm for different combination of training and test set. Moreover, we also have compared the result of our method with the other competent methods. In general the face images are severely affected by the different environmental condition. These factors need to be investigated to measure their impact on the intra-class assignment. The scatter matrices involve the overlapping sample distribution information for classification.

In this experiment, UMIST database, to generate distinct pair of the training and test sets we have taken the s as 4, 6, 8 and 10. In this context, each pair of training and test sets is disjoint in nature. The performance of the proposed technique is performed by considering each value of s with 20 dissimilar pairs of training and test sets on the UMIST face database. Fig. 4 also shows the minimum, maximum and average recognition rates (%) of the FG-2DLDA method for different values of s by varying feature size. Table 4 shows a comparative presentation of the FG-2DLDA method along with other contemporary methods in terms of best average recognition rates. The proposed method yields the best average recognition rates (dimension of feature vector) of 86.81% (18×18), 92.75% (20×20), 96.83% (14×14), and 97.3% (14×14) for s = 4, 6, 8 and 10, respectively. In this case, the discriminative information is extracted by calculating fuzzy scatter matrices. The discriminative projection vectors are obtained when the fuzzy scatter matrices are singular. The results show that in all the cases, the performance of the FG-2DLDA method is superior to the other methods.

5 Conclusion

In this paper, fuzzy generalized two-dimensional Fisher's linear discriminant analysis (FG-2DLDA) method for face recognition is presented. This method assumes that a face image may belong to several classes with possibility of different membership values. These membership values are generated by fuzzy k-NN algorithm and used to generate fuzzy global mean image and fuzzy class-wise mean images. Finally these mean images are used to generate fuzzy intra-class and inter-class scatter matrices along row and column directions. The projection matrices obtained by solving these scatter matrices, satisfying the two Fisher's criteria, yield rich information leading to generation of superior discriminant features. Image classification and recognition is performed using a RBF neural network. The performance of our method is validated on the FERET, AT&T and UMIST and face databases. The experimental results demonstrate that the FG-2DLDA method outperforms the competent methods.

Acknowledgement

This work was supported by the Senior Research fellowship Program of Aniruddha Dey under the State Government Fellowship (Ref. No. - P-1/RS./365/12, dated 05th October, 2012.) Of the Department of Computer Science & Engineering, Jadavpur University, Kolkata.

References

[1] R. Chellappa, C. L. Wilson, and S. Sirohey. (1995) Human and machine recognition of faces: a survey. *Proc. IEEE* vol. 83, 705–740. https://doi.org/10.1109/5.381842

[2] W. Zhao, R. Chellappa, and P. J. Phillops. (2003) A. Rosenfeld. Face recognition: a literature survey. *ACM Comput.* Surveys. 35: 399–458. https://doi.org/10.1145/954339.954342

[3] A. S. Tolba, A.H. El-Baz, and A.A. El-Harby. (2006) Face recognition: a literature review. *Int. J. Signal Process*, 2: 88–103.

[4] H. Zhou, A. Mian, L. Wei, D. Creighton, M. Hossny, and S. Nahavandi. (2014) Recent advances on singlemodal and multimodal face recognition: a survey. *IEEE Trans. Human Machine Systems*, 44(6): 701–716. https://doi.org/10.1109/THMS.2014.2340578

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. (1997) Eigenfaces versus fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19:711–720. https://doi.org/10.1109/34.598228

[6] B. Poon, M. A. Amin, and H. Yan. (2011) Performance evaluation and comparison of PCA Based human face recognition methods for distorted images. International Journal of Machine Learning and Cybernetics, 2(4): 245-259. https://doi.org/10.1007/s13042-011-0023-2

[7] G. J. Alvarado, W. Pedrycz,·M. Reformat, and K.-C. Kwak. (2006) Deterioration of visual information in face classification using eigenfaces and fisherfaces. *Machine Vision and Applications*, 17(1): 68–82. https://doi.org/10.1007/s00138-006-0016-4

[8] L. F Chen, H. Y Mark Liao, M. T. Ko, J.C Lin, and G. J.Yu. (2000) A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recogn.*, 33: 1713–26. https://doi.org/10.1016/S0031-3203(99)00139-9

[9] [9] H. Yu, and J. Yang. (2001) A direct LDA algorithm for high-dimensional data-with application to face recognition. *Pattern Recogn*. 34: 2067–70. https://doi.org/10.1016/S0031-3203(00)00162-X

[10] X. S. Zhuang, and D. Q. Dai. (2007) Improved discriminant analysis for high-dimensional data and its application to face recognition. *Pattern Recogn.*, 40(5): 1570-1578. https://doi.org/10.1016/j.patcog.2006.11.015

[11] D. Swets, and J. Weng. (1996) Using discriminant eigenfeatures for image retrieval. *IEEE Trans. Pattern Anal. Machine Intell.*, 18(8) 831–836. https://doi.org/10.1109/34.531802

[12] J. Friedman. (1989) Regularized discriminant analysis. *J. Am. Stat. Assoc.*, 165– 175. https://doi.org/10.1080/01621459.1989.10478752

[13] R. Huang, Q. Liu, H. Lu, and S. Ma. (2002) Solving the small sample size problem of lda. *In Proceedings of the 16th International Conference on Pattern Recognition*, 3, 29–32. https://doi.org/10.1109/ICPR.2002.1047787

[14] H. Yu, and J. Yang. (2001) A direct lda algorithm for high-dimensional data-with application to face recognition. *Pattern Recogn.*, 34 (10): 2067- 2075. https://doi.org/10.1016/S0031-3203(00)00162-X

[15] M. S. Bartlet, J. R. Movellan, and T. J. Sejnowski. (2002) Face recognition by independent component analysis. *IEEE Trans. Neural Netw.*, 13(6):1450-1464. https://doi.org/10.1109/TNN.2002.804287

[16] J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang. (2004) Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(1):131–137. https://doi.org/10.1109/TPAMI.2004.1261097

[17] H. Xiong, M. N. S. Swamy, and M. O. Ahmad. (2005) Two-dimensional FLD for face recognition. *Pattern Recogn.*, 38(7):1121–1124. https://doi.org/10.1016/j.patcog.2004.12.003

[18] S. Chowdhury, J. K. Sing, D. K. Basu, and M. Nasipuri. (2011) Face recognition by generalized two-dimensional FLD method and multi-class support vector machines. *Appl. Soft Comput.*,11(7):4282–4292. https://doi.org/10.1016/j.asoc.2010.12.002

[19] H. Yu, and J. Yang. (2001) A direct LDA algorithm for high-dimensional data with application to face recognition, Pattern Recogn., 34: 2067–2070. https://doi.org/10.1016/S0031-3203(00)00162-X

[20] J. Yang, and J. Yang. (2003) Why can LDA be performed in PCA transformed space? Pattern Recognit. 36 (2): 563-566. https://doi.org/10.1016/S0031-3203(02)00048-1

[21] J. Ye, and Q. Li. (2004) LDA/QR: An efficient and effective dimension reduction algorithm and its theoretical foundation, Pattern Recognit. 37 (4), 851–854. https://doi.org/10.1016/j.patcog.2003.08.006

[22] J. Ye, R. Janardan, C. Park, and H. Park. (2004) An optimization criterion for generalized discriminant analysis on undersampled problems, IEEE Trans. Pattern Anal. Mach. Intell. 26 (8): 982–994. https://doi.org/10.1109/TPAMI.2004.37

[23] J. M. Keller, M. R. Gray, and J. A. Givens. (1985) A fuzzy *k*-nearest neighbor algorithm. *IEEE Trans. Syst. Man Cybernet.*, 15(4):580–585. https://doi.org/10.1109/TSMC.1985.6313426

[24] K. C. Kwak, and W. Pedrycz. (2005) Face recognition using a fuzzy fisherface classifier.

Journal of the Pattern Recogn. 38(10):1717-1732. https://doi.org/10.1016/j.patcog.2005.01.018

[25] W. Yang, J. Wang, M. Ren, and J. Yang. (2009) Fuzzy 2-dimensional FLD for face recognition. *Journal of Information and Computing Science*, 4(3): 233-239. https://doi.org/10.1109/CCPR.2009.5344077

[26] W. Yang, J. Wang, M. Ren, L. Zhang, and J. Yang. (2009) Feature extraction using fuzzy inverse FDA. *Neurocomputing*, 72(13- 15): 3384–3390. https://doi.org/10.1016/j.neucom.2009.03.011

[27] X. N.Song, Y. J. Zheng, X. J.Wu, X. B.Yang, and J. Y.Yang. (2010) A complete fuzzy discriminant analysis approach for face recognition. *Applied Soft Computing*, 10: 208-214. https://doi.org/10.1016/j.asoc.2009.07.002

[28] L. Xiaodong, F. Shumin, and T. Zhang. (2013) Weighted maximum scatter difference based feature extraction and its application to face recognition. *Machine Vision and Applications*, 22: 591-595.

[29] M. Zhao, T. W. S. Chow, and Z. Zhang (2012) Random walk-based fuzzy linear discriminant analysis for dimensionality reduction. *Soft Computing*, 16:1393-1409. https://doi.org/10.1007/s00500-012-0843-3

[30] J. Wang, W. Yang, and J. Yang. (2013) Face recognition using fuzzy maximum scatter discriminant analysis. *Neural Computing & Application*, 23: 957-964. https://doi.org/10.1007/s00521-012-1020-4

[31] X. Li, and A. Song. (2013) Fuzzy MSD based feature extraction method for extraction. *Neurocomputing*, 122: 266-271. https://doi.org/10.1016/j.neucom.2013.06.025

[32] X. Li (2014) Face recognition method based on fuzzy 2DPCA. *Journal of Electrical and Computer Engineering*, 2014:1- 7. https://doi.org/10.1155/2014/919041

[33] N. Zheng, L. Qi, and L. Guan. (2014) Generalised multiple maximum scatter difference feature extraction using QR decomposition. *Journal of visual Communication Image Representation*, 25:1460-1471. https://doi.org/10.1016/j.jvcir.2014.04.009

[34] J.K. Sing. (2015) A novel Gaussian probabilistic generalized 2DLDA for feature extraction and face recognition. *In Proceedings of the IEEE Conference on Computer Graphics, Vision and Information Security*, pages 258-263. https://doi.org/10.1109/CGVIS.2015.7449933

[35] P. Huang, Z. Yang, and C. Chen. (2015) Fuzzy local discriminant embedding for image feature extraction. *Computers and Electrical Engineering*, 46:231-240. https://doi.org/10.1016/j.compeleceng.2015.03.013

[36] J. Xu, Z. Gu, and K. Xie. Fuzzy local mean discriminant analysis for dimensionality reduction. *Neural Processing Letter*, 44:701-718, 2016. https://doi.org/10.1007/s11063-015-9489-3

[37] P. Huang, G. Gao, C. Qian, G. Yang, and Z. Yang. (2017) Fuzzy linear regression discriminant projection for face recognition. *IEEE Access*, 23:169-174. https://doi.org/10.1109/ACCESS.2017.2680437

[38] Q. Zhu**,** and Y. Xu. (2013) Multi-directional two dimensional PCA with matching score level fusion for face recognition. *Neural Comput. & Applic.*, 23(1): 169-174. https://doi.org/10.1007/s00521-012-0851-3

[39] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. (2000) The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern. Anal. Mach. Intell.*, 22: 1090–1104. https://doi.org/10.1109/34.879790

[40] P. J. Phillips. (2004) The Facial Recognition Technology (FERET) database. <http://www.itl.nist.gov/iad/humanid/feret/feret_master.html>.

[41] The ORL face database, <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>.

[42] D. B. Graham, N. M. Allinson, H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soulie, and T. S. Huang (Eds.), (1998) Characterizing virtual eigen signatures for general purpose face recognition: From theory to applications. *NATO ASI Series F Computer and Systems Sciences*, 163: 446–456. https://doi.org/10.1007/978-3-642-72201-1_25

https://doi.org/10.31449/inf.v43i4.3007 Informatica **43** (2019) 545–550 **545**

The MAP/G/1 G-queue with Unreliable Server and Multiple Vacations

Yi Peng
School of Mathematical Science, Changsha Normal University, Changsha 410100, Hunan, P.R. China
E-mail: scgyp06@163.com

Keywords: G-queues, Markovian Arrival Process (MAP), reliability, censoring technique, RG-factorization

Received: November 21, 2018

In this paper, we consider a MAP/G/1 G-queues with unreliable server and multiple vacations. The arrival of a negative customer not only removes the customer being in service, but also makes the server under repair. The server leaves for a vacation as soon as the system empties and is allowed to take repeated (multiple) vacations. By using the supplementary variables method and the censoring technique, we obtain the queue length distributions. We derive the mean of the busy period based on the renewal theory. Furthermore, we analyze some main reliability indexes and investigate some important special cases.

Povzetek: Predstavljena je nova metoda obravnave strežniških vrst v pogojih nezanesljivega delovanja in v primerih občasne odsotnosti obdelave.

1 Introduction

Recently there has been a rapid increase in the literature on queueing systems with negative arrivals. Queues with negative arrivals, called G-queues, were first introduced by Gelenbe [1]. When a negative customer arrives at the queue, it immediately removes one or more positive customers if present. Negative arrivals have been interpreted as viruses, orders of demand, inhibiter. Queueing systems with negative arrivals have many applications in computer, neural networks, manufacturing systems and communication networks etc. There is a lot of research on queueing system with negatives arrivals. For a comprehensive survey on queueing systems with negative arrivals, readers may see [1-4].

Boucherie and Boxma [3] considered an M/G/1 queue with negative arrivals where a negative arrival removes a random amount of work. Li and Zhao [5] discussed an MAP/G/1 queue with negative arrivals. They analyzed two classes of removal rules: (i) arrival of a negative customer which removes all the customers in the system (RCA); (ii) arrival of a negative customer which removes only a customer from the head of the system (RCH), including the customer being in service.

Queueing system with repairable server has been studied by many authors such as Cao and Chen [6], Neuts and Lucantoni [7]. Wang, Cao and Li [8] analyzed the reliability of the retrial queues with server breakdowns and repairs. Harrison and Pitel [9]considered the M/M/1 G-queues with breakdowns and exponential repair times. Li, Ying and Zhao [10] investigated a BMAP/G/1 retrial queue with a server subject to breakdowns and repairs.

For a detailed survey on queueing systems with server vacations one can refer to Refs [11]. Recently, Sikdar and Gupta [12] discussed the queue length distributions in the finite buffer bulk-service MAP/G/1 queue with multiple vacations. Kasahara, Takine, Takahashi and Hasegawa [13] considered the MAP/G/1 queues under N-policy with and without vacations.

Most of the analysis in the past have been carried out assuming Poisson input. However, in recent years there has been a growing interest to analyze queues by considering input process as Markovian arrival process (MAP). The MAP is a useful mathematical model for describing bursty traffic in modern communication networks, and is a rich class of point processes containing many familiar arrival processes such as Poission process, PH-renewal process, Markov modulated Poission process, etc. Readers may refer to chapter 8 in Bocharov [14].

In this paper, we consider the MAP/G/1 G-queues with unreliable server and multiple vacations. The process of arrivals of negative customers is also MAP. The arrival of a negative customer not only removes the customer being in service, but also makes the server under repair. We obtain the distributions of stationary queue length, the mean of the busy period and some reliability indexes by using the supplementary variable method, the matrix-analytic method, the censoring technique, and the renewal theory.

The rest of this paper is organized as follows. The model description is given in section 2. The stationary differential equations of the model and their solutions are obtained in section 3. The expressions for the distributions of the stationary queue length and the mean of the busy period are derived in section 4. Some special cases are considered in section 5. Some numerical examples are shown in section 6.

2 Model description

In this section, we consider a single server queue with two types of independent arrivals, positive and negative. Positive arrivals correspond to customers who upon arrival, join the queue with the intention of being served and then leaving the system. At a negative arrival epoch, the system is affected if and only if the server is working.

The arrival process. We assume that the arrivals of both positive and negative customers are MAPs with matrix descriptors (C_1, D_1) and (C_2, D_2) respectively, where the infinitesimal generators $C_1 + D_1$ and $C_2 + D_2$ of sizes $m_1 \times m_1$ and $m_2 \times m_2$, respectively, are irreducible and positive recurrent. Let θ_1 and θ_2 be the stationary probability vectors of $C_1 + D_1$ and $C_2 + D_2$, respectively. Then $\lambda_1 = \theta_1 D_1 e$ and $\lambda_2 = \theta_2 D_2 e$ are the stationary arrival rates of positive and negative customers, respectively, where is a column vector of ones of a suitable size.

The removal rule. The arrival of a negative customer not only removes the customer being in service, but also makes the server under repair. And after repair the server is as good as new. As soon as the repair of the server is completed, the server enters the working state immediately and continues to serve the next customer if the queue is not empty.

The vacations. When the server finishes serving a positive customer or the repair of the server is completed and finds the queue empty, the server leaves for a vacation of random length V. On return from a vacation if he finds more than one customer waiting, he takes the customer from the head of the queue for service and continues to serve in this manner until the queue is empty. Otherwise, he immediately goes for another vacation.

The service time. All positive customers have i.i.d. service time distribution given by

[Beginning of the document] [Automatic section break]

$$B_S(x) = 1 - \exp\left\{-\int_0^x \mu_S(\nu)\, d\nu\right\} \quad \text{with mean } 1/\mu_S \in (0, +\infty).$$

The vacation time. The vacation time distribution is given by

$$B_V(x) = 1 - \exp\left\{-\int_0^x \mu_V(\nu)\, d\nu\right\} \quad \text{with mean } 1/\mu_V \in (0, +\infty).$$

The repair time. The repair time distribution is given by

$$B_R(x) = 1 - \exp\left\{-\int_0^x \mu_R(\nu)\, d\nu\right\}$$

with mean $1/\mu_R \in (0, +\infty)$.The independence. We assume that all the random variables defined above are independent. Throughout the rest of the paper, we denote by $\overline{F}(x) = 1 - F(x)$ the tail of distribution function $F(x)$.

3 The differential equations and the solution

In this section, we first introduce several supplementary variables to construct the differential equations for the model. We then use the censoring technique to solve these equations. The solution to the differential equations will be used to obtain interesting performance measures of the system in later sections.

Let $N(t)$ be the number of customers in the system at time t , and let $J_1(t)$ and $J_2(t)$ be the phases of the arrivals of positive and negative customers at time t, respectively. We define the states of the server as

$$I(t) = \begin{cases} S, & \text{if the server is working with service time distribution } B_S(x), \\ V, & \text{if the server is on vacation with vacation time distribution } B_V(x), \\ R, & \text{if the server is being repaired with repair time distribution } B_R(x). \end{cases}$$

For $t > 0$, we define the random variable $S(t)$ as follows:(i) if $I(t) = S$, $S(t)$ represents the elapsed service time received by a customer with the service time up to time t; (ii)if $I(t) = V$, $S(t)$ represents the elapsed vacation time up to time t; (iii) if $I(t) = R$, $S(t)$ represents the elapsed repair time up to time t. Then, $\{I(t), N(t), J_1(t), J_2(t), S(t) : t \geq 0\}$ is a Markov process. The state space of the process is expressed as

$$\begin{aligned} \Omega = & \{(S, k, j_1, j_2, x) : k \geq 1, 1 \leq j_1 \leq m_1, 1 \leq j_2 \leq m_2, x \geq 0\} \\ & \cup \{(V, k, j_1, j_2, x) : k \geq 0, 1 \leq j_1 \leq m_1, 1 \leq j_2 \leq m_2, x \geq 0\} \\ & \cup \{(R, k, j_1, j_2, x) : k \geq 0, 1 \leq j_1 \leq m_1, 1 \leq j_2 \leq m_2, x \geq 0\} \end{aligned}$$

We write:

$$p_{S,k,i,j}(t,x)dx = p\{I(t) = S, N(t) = k, J_1(t) = i, J_2(t) = j, x \leq S(t) < x + dx\},$$
$$p_{V,k,i,j}(t,x)dx = p\{I(t) = V, N(t) = k, J_1(t) = i, J_2(t) = j, x \leq S(t) < x + dx\},$$
$$p_{R,k,i,j}(t,x)dx = p\{I(t) = R, N(t) = k, J_1(t) = i, J_2(t) = j, x \leq S(t) < x + dx\},$$
$$p_{S,k,i,j}(x) = \lim_{t \to +\infty} p_{S,k,i,j}(t,x),$$
$$p_{V,k,i,j}(x) = \lim_{t \to +\infty} p_{V,k,i,j}(t,x),$$
$$p_{R,k,i,j}(x) = \lim_{t \to +\infty} p_{R,k,i,j}(t,x),$$
$$P_k^S(x) = (p_{S,k,1,1}(x), \cdots, p_{S,k,1,m_2}(x), \cdots, p_{S,k,m_1,1}(x), \cdots, p_{S,k,m_1,m_2}(x)),$$
$$P_k^V(x) = (p_{V,k,1,1}(x), \cdots, p_{V,k,1,m_2}(x), \cdots, p_{V,k,m_1,1}(x), \cdots, p_{V,k,m_1,m_2}(x)),$$
$$P_k^R(x) = (p_{R,k,1,1}(x), \cdots, p_{R,k,1,m_2}(x), \cdots, p_{R,k,m_1,1}(x), \cdots, p_{R,k,m_1,m_2}(x)).$$

If the system is stable, then the system of stationary differential equations of the joint probability density $\{P_0^V(x), P_0^R(x), P_k^S(x), P_k^V(x), P_k^R(x), k \geq 1\}$ can be written as

$$\frac{d}{dx}P_1^S(x) = P_1^S(x)[C_1 \oplus C_2 - \mu_S(x)I], \quad (1)$$
$$\frac{d}{dx}P_k^S(x) = P_k^S(x)[C_1 \oplus C_2 - \mu_S(x)I] + P_{k-1}^S(x)(D_1 \otimes I), k \geq 2, \quad (2)$$
$$\frac{d}{dx}P_0^V(x) = P_0^V(x)[C_1 \oplus C_2 + I \otimes D_2 - \mu_V(x)I], \quad (3)$$
$$\frac{d}{dx}P_k^V(x) = P_k^V(x)[C_1 \oplus C_2 + I \otimes D_2 - \mu_V(x)I] + P_{k-1}^V(x)(D_1 \otimes I), k \geq 1, \quad (4)$$

$$\frac{d}{dx}P_0^R(x) = P_0^R(x)[C_1 \oplus C_2 + I \otimes D_2 - \mu_R(x)I], \quad (5)$$

$$\frac{d}{dx}P_k^R(x) = P_k^R(x)[C_1 \oplus C_2 + I \otimes D_2 - \mu_R(x)I] + P_{k-1}^R(x)(D_1 \otimes I), k \geq 1. \quad (6)$$

The joint probability density $\{P_0^V(x), P_0^R(x), P_k^S(x), P_k^V(x), P_k^R(x), k \geq 1\}$ should satisfy the boundary conditions:

$$P_0^V(0) = \int_0^{+\infty} P_1^S(x)\mu_S(x)\,dx + \int_0^{+\infty} P_0^V(x)\mu_V(x)\,dx + \int_0^{+\infty} P_0^R(x)\mu_R(x)\,dx, \quad (7)$$

$$P_k^V(0) = 0, \quad k \geq 1, \quad (8)$$

$$P_k^R(0) = \int_0^{+\infty} P_{k+1}^S(x)\,dx(I \otimes D_2), \quad k \geq 0, \quad (8)$$

$$P_k^S(0) = \int_0^{+\infty} P_{k+1}^S(x)\mu_S(x)\,dx + \int_0^{+\infty} P_k^V(x)\mu_V(x)\,dx + \int_0^{+\infty} P_k^R(x)\mu_R(x)\,dx \quad k \geq 1, \quad (9)$$

and the normalization condition:

$$\left\{\sum_{k=0}^{\infty}\int_0^{+\infty} P_k^V(x)\,dx + \sum_{k=0}^{\infty}\int_0^{+\infty} P_k^R(x)\,dx + \sum_{k=1}^{\infty}\int_0^{+\infty} P_k^S(x)\,dx\right\} e = 1. \quad (11)$$

In the remainder of this section, we solve equations (1)-(11). To sovle equations (1)-(6). we define $Q_S^*(z,x) = \sum_{k=1}^{\infty} z^k P_k^S(x),\ Q_V^*(z,x) = \sum_{k=0}^{\infty} z^k P_k^V(x),\ Q_R^*(z,x) = \sum_{k=0}^{\infty} z^k P_k^R(x).$

It follows from (1) and (2) that

$$\frac{\partial}{\partial x}Q_S^*(z,x) = Q_S^*(z,x)[(C_1 + zD_1) \oplus C_2 - \mu_S(x)I],$$

which leads to

$$Q_S^*(z,x) = Q_S^*(z,0)[\exp\{(C_1 + zD_1)x\} \otimes \exp\{C_2x\}]\overline{B_S}(x). \quad (12)$$

It follows from (3) and (4) that

$$\frac{\partial}{\partial x}Q_V^*(z,x) = Q_V^*(z,x)[(C_1 + zD_1) \oplus (C_2 + D_2) - \mu_V(x)I],$$

which leads to

$$Q_V^*(z,x) = Q_V^*(z,0)[\exp\{(C_1 + zD_1)x\} \otimes \exp\{(C_2 + D_2)x\}]\overline{B_V}(x). \quad (13)$$

It follows from (5) and (6) that

$$\frac{\partial}{\partial x}Q_R^*(z,x) = Q_R^*(z,x)[(C_1 + zD_1) \oplus (C_2 + D_2) - \mu_R(x)I],$$

which leads to

$$Q_R^*(z,x) = Q_R^*(z,0)[\exp\{(C_1 + zD_1)x\} \otimes \exp\{(C_2 + D_2)x\}]\overline{B_R}(x). \quad (14)$$

Let us define $P(n,t), n \geq 0, t \geq 0$ as $m_1 \times m_1$ matrix whose element $(P(n,t))_{ij}$ is the probability that exactly n positive customers arrive during $[0,t)$ and the generation process passes from phase i to phase j. These matrices satisfy the following system of differential equations

$$\begin{aligned}\frac{d}{dt}P(0,t) &= P(0,t)C_1,\\ \frac{d}{dt}P(n,t) &= P(n,t)C_1 + P(n-1,t)D_1, \quad n \geq 1,\end{aligned}$$

with $P(0,0) = I$.We define

$$P^*(z,t) = \sum_{n=0}^{\infty} z^n P(n,t), \quad |z| \leq 1,$$

Solving the above matrix differential equation,we get

$$P^*(z,t) = e^{(C_1+zD_1)t}, \quad |z| \leq 1, t \geq 0. \quad (15)$$

Substituting (15) into (12)-(14) respectively gives

$$P_k^S(x) = \sum_{j=1}^{k} P_j^S(0)[P(k-j,x) \otimes \exp\{C_2x\}]\overline{B_S}(x), k \geq 1, \quad (16)$$

$$\begin{aligned}P_k^V(x) &= \sum_{j=0}^{k} P_j^V(0)[P(k-j,x) \otimes \exp\{(C_2+D_2)x\}]\overline{B_V}(x)\\ &= P_0^V(0)[P(k,x) \otimes \exp\{(C_2+D_2)x\}]\overline{B_V}(x), \quad k \geq 0. \quad (17)\end{aligned}$$

$$P_k^R(x) = \sum_{j=0}^{k} P_j^R(0)[P(k-j,x) \otimes \exp\{(C_2+D_2)x\}]\overline{B_R}(x), k \geq 0, \quad (18)$$

Equations (16)-(18) provide a solution for the system of differential equations (1)-(6). Furthermore, boundary equations (7)-(10) will be used to determine the vectors $P_k^S(0)$ for $k \geq 1$, $P_k^R(0)$ for $k \geq 0$ and $P_k^V(0)$ for $k \geq 0$.We define:

$$A_k = \int_0^{+\infty} [P(k,x) \otimes \exp\{C_2x\}]\,dB_S(x),$$

$$B_k = \int_0^{+\infty} [P(k,x) \otimes \exp\{(C_2+D_2)x\}]\,dB_R(x)$$

$$W_k = \int_0^{+\infty} [P(k,x) \otimes \exp\{(C_2+D_2)x\}]\,dB_V(x),$$

$$E_k = \int_0^{+\infty} [P(k,x) \otimes \exp\{C_2x\}]\overline{B_S}(x)\,dx(I \otimes D_2)$$

Then it follows from (7)-(10),(16)-(18)that $P = P\Pi$, where

$$P = (P_0^V(0), P_1^S(0), P_0^R(0), P_2^S(0), P_1^R(0), \cdots) \quad (19)$$

and

$$\Pi = \begin{pmatrix} W_0 & \widetilde{W_1} & \widetilde{W_2} & \widetilde{W_3} & \cdots\\ H_0 & \widetilde{A_1} & \widetilde{A_2} & \widetilde{A_3} & \cdots\\ & \widetilde{A_0} & \widetilde{A_1} & \widetilde{A_2} & \cdots\\ & & \widetilde{A_0} & \widetilde{A_1} & \cdots\\ & & & \widetilde{A_0} & \cdots\\ & & & & \ddots \end{pmatrix},$$

$$\widetilde{W_k} = (W_k, 0), \quad H_0 = \begin{pmatrix} A_0\\ B_0 \end{pmatrix}$$

With

$$\widetilde{A_k} = \begin{pmatrix} A_k & E_{k-1}\\ B_k & 0 \end{pmatrix}, \quad E_{-1} = 0, \quad k \geq 0,$$

Therefore, we obtain the transition probability matrix and stationary differential equations of the system.

4 Performance measures of the model

In this section,we consider two performance measures for the model: the stationary queue length, the busy period.

4.1 The stationary queue length

We write

$$\begin{aligned}p_k &= \lim_{t\to\infty} P\{N(t) = k\}, \quad k \geq 0,\\ p_k^S &= \lim_{t\to\infty} P\{N(t) = k, I(t) = S\}, \quad k \geq 1,\\ p_k^V &= \lim_{t\to\infty} P\{N(t) = k, I(t) = V\}, \quad k \geq 0,\\ p_k^R &= \lim_{t\to\infty} P\{N(t) = k, I(t) = R\}, \quad k \geq 0.\end{aligned}$$

Obviously,

$$p_0 = p_0^V + p_0^R\ ; \quad p_k = p_k^S + p_k^V + p_k^R, \quad k \geq 1.$$

Theorem 1 If the model is stable, then

$$\begin{cases} p_0 = \beta(x_0 H_0^V + x_0^R H_0^R)e, \\ p_k = \beta x_0 H_k^V e + \beta \sum_{j=1}^{k} x_j^S H_{k-j}^S e + \beta \sum_{j=0}^{k} x_j^R H_{k-j}^R e, \quad k \geq 1, \end{cases}$$

where

$$H_k^S = \int_0^{+\infty} [P(k,x) \otimes \exp\{C_2 x\}] \overline{B_S}(x)\, dx, \quad k \geq 0,$$

$$H_k^V = \int_0^{+\infty} [P(k,x) \otimes \exp\{(C_2 + D_2)x\}] \overline{B_V}(x)\, dx, \quad k \geq 0,$$

$$H_k^R = \int_0^{+\infty} [P(k,x) \otimes \exp\{(C_2 + D_2)x\}] \overline{B_R}(x)\, dx, \quad k \geq 0,$$

So that the mean number of customers in the system is

$$L = \sum_{k=0}^{\infty} k p_k = \beta x_0 \sum_{k=1}^{\infty} k H_k^V e + \beta \sum_{k=1}^{\infty} k \sum_{j=1}^{k} x_j^S H_{k-j}^S e + \beta \sum_{k=1}^{\infty} k \sum_{j=0}^{k} x_j^R H_{k-j}^R e.$$

Proof: It follows from (17) and (18) that

$$p_0 = \int_0^{+\infty} [P_0^V(x) + P_0^R(x)]\, dx\ e = \beta(x_0 H_0^V + x_0^R H_0^R)e,$$

and from (16)-(18) that

$$p_k^S = \int_0^{+\infty} P_k^S(x)\, dx\ e = \beta \sum_{j=1}^{k} x_j^S H_{k-j}^S e,$$

$$p_k^V = \int_0^{+\infty} P_k^V(x)\, dx\ e = \beta x_0 H_k^V e,$$

$$p_k^R = \int_0^{+\infty} P_k^R(x)\, dx\ e = \beta \sum_{j=0}^{k} x_j^R H_{k-j}^R e.$$

This completes the proof.

4.2 The busy period

We now provide an analysis of the busy period (including of the period when the server is under repair) of the model.

Let V be the random variable of the vacation time, or $B_V(x) = P\{V \leq x\}$.

We denote by T be the random variable of the interarrival time between two positive customers, and $T^{(E)}$ the random variable for the equilibrium excess distributions with respect to T. Then we have

$$A(x) = P\{T \leq x\} = \theta_1 \int_0^x \exp\{C_1 t\}\, dt\ D_1 e$$

And

$$A^{(E)}(x) = P\{T^{(E)} \leq x\} = \frac{1}{\theta_1(-C_1)^{-1}e} \int_0^x \overline{A}(t)\, dt.$$

Let V_i be the random variable of the i-th vacation, and $\hat{V}$ be the random variable of the number of times of vacations during the total vacation period. Then

$$\begin{aligned} P\{\hat{V} = n\} &= P\{\sum_{i=1}^{n-1} V_i < T^{(E)} \leq \sum_{i=1}^{n} V_i\} \\ &= \int_0^{+\infty} [B_V^{(n-1)*}(t) - B_V^{n*}(t)]\, dA^{(E)}(t), \end{aligned}$$

We denote by $F(x) * G(x)$ the convolution of two functions $F(x)$ and $G(x)$ given by $F(x) * G(x) = \int_0^x F(x-u)\, dG(u)$. We write $F^{n*}(x) = F(x) * F^{(n-1)*}(x)$ for $n \geq 2$ and define $F^{0*}(x) = 1$.

Lemma 1 Let $\overline{V}$ be the random variable of the length of multiple vacations, then

$$E\overline{V} = \sum_{n=1}^{\infty} n \frac{1}{\mu_V} \int_0^{+\infty} [B_V^{(n-1)*}(t) - B_V^{n*}(t)]\, dA^{(E)}(t).$$

Theorem 2 Let ξ be the random variable of the busy period of the system, then

$$E\xi = \frac{(1 - \beta x_0 L_V e) E\overline{V}}{\beta x_0 L_V e}.$$

Proof: According to the renewal theory, we can obtain

$$\sum_{k=0}^{\infty} p_k^V = \frac{E\overline{V}}{E\xi + E\overline{V}},$$

or

$$E\xi = \frac{(1 - \sum_{k=0}^{\infty} p_k^V) E\overline{V}}{\sum_{k=0}^{\infty} p_k^V} = \frac{(1 - \beta x_0 L_V e) E\overline{V}}{\beta x_0 L_V e}.$$

This completes the proof.

Consequently, we obtain some important performance measures for the model: the stationary queue length, the mean number of customers in the system, the mean length of multiple vacations and the mean busy period.

5 Special cases

In this section we will investigate very briefly some important special cases.

Case 1. No negative arrival takes place and the server is reliable.

In this case, our model becomes the MAP/G/1 queue with multiple vacations.

We put $C_2 = D_2 = 0$ and $B_R(x) = 0$ in the main results and obtain

$$Q = \begin{pmatrix} A_1 & A_2 & A_3 & \cdots \\ A_0 & A_1 & A_2 & \cdots \\ & A_0 & A_1 & \cdots \\ & & A_0 & \cdots \\ & & & \ddots \end{pmatrix},$$

$$A_k = \int_0^{+\infty} P(k,x)\, dB_S(x), \quad W_k = \int_0^{+\infty} P(k,x)\, dB_V(x),$$

$$R_j = \sum_{i=1}^{\infty} A_{i+j} G_1^{i-1} [I - \Phi_0]^{-1}, j \geq 1, \quad R_{0,j} = \sum_{i=0}^{\infty} W_{i+j} G_1^i [I - \Phi_0]^{-1}, j \geq 1,$$

$$\Psi_0 = W_0 + \sum_{i=0}^{\infty} W_{i+1} G_1^i [I - \sum_{i=1}^{\infty} A_i G_1^{i-1}]^{-1} A_0, \quad G_1 = \hat{Q}(1,1) A_0,$$

$$L_V = \int_0^{+\infty} \exp\{(C_1 + D_1)x\} \overline{B_V}(x)\, dx, \quad L_S = \int_0^{+\infty} \exp\{(C_1 + D_1)x\} \overline{B_S}(x)\, dx,$$

$$\begin{cases} p_0 = \beta x_0 H_0^V e, \\ p_k = \beta x_0 H_k^V e + \beta \sum_{j=1}^{k} x_j^S H_{k-j}^S e, \quad k \geq 1, \end{cases}$$

$$H_k^S = \int_0^{+\infty} P(k,x) \overline{B_S}(x)\, dx, \quad H_k^V = \int_0^{+\infty} P(k,x) \overline{B_V}(x)\, dx, \quad k \geq 0.$$

Case 2. No vacation is allowed, in this case, our model becomes the MAP/G/1 G-queue with unreliable server.We assume that $B_V(x) = 0$ in the main results and obtain

$$W_0 = 0, \quad \widetilde{W_1} = D_1 \otimes I, \quad \widetilde{W_k} = 0, \quad k \geq 2,$$

$$R_{0,1} = \widetilde{W_1}[I - \Phi_0]^{-1}, \quad R_{0,j} = 0, \quad j \geq 2,$$

$$\beta = \frac{1}{x_0 e + \sum\limits_{k=1}^{\infty} x_k^S L_S e + \sum\limits_{k=0}^{\infty} x_k^R L_R e}, \quad \Psi_0 = \widetilde{W_1}[I - \sum_{i=1}^{\infty} \widetilde{A}_i G_1^{i-1}]^{-1} H_0,$$

$$\begin{cases} p_0 = \beta(x_0 + x_0^R H_0^R)e, \\ p_k = \beta x_0 e + \beta \sum_{j=1}^{k} x_j^S H_{k-j}^S e + \beta \sum_{j=0}^{k} x_j^R H_{k-j}^R e, \quad k \geq 1. \end{cases}$$

We note that these results are consistent with the known results in [5] and [13].

6 Numerical examples

In this section, we discuss some interesting numerical examples that qualitatively describe the performance of the queueing model under study. The following examples are illustrated using the results of section 3. The algorithms have been written into a MATLAB program. For the purpose of a numerical illustration, we assume that all distribution functions in this paper are exponential, i.e. $B_S(x), B_V(x), B_R(x)$ are exponential distribution functions and their parameters are $\mu_S = 2.158, \mu_V, \mu_R$ respectively. Also, we vary values of μ_V, μ_R such that the system is stable. Numerical results are presented in Figures 1-4.

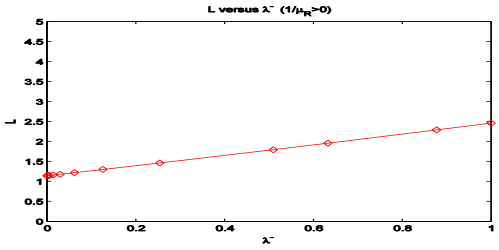


Figure 1: The mean system size versus λ^- with $(\mu_V, \mu_R) = (1.603, 0.911)$.

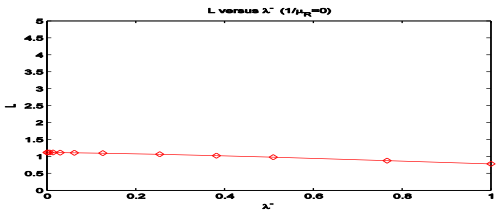


Figure 2: The mean system size versus λ^- with $(\mu_V, 1/\mu_R) = (1.603, 0)$.

Here we choose the following arbitrary values:

$$C_1 = \begin{pmatrix} -0.7 & 0.2 \\ 0.3 & -1.4 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0.4 & 0.1 \\ 0.2 & 0.9 \end{pmatrix},$$

$$C_2 = \lambda^- \begin{pmatrix} -0.007 & 0.002 \\ 0.03 & -0.009 \end{pmatrix}, \quad D_2 = \lambda^- \begin{pmatrix} 0.001 & 0.004 \\ 0.005 & 0.001 \end{pmatrix}.$$

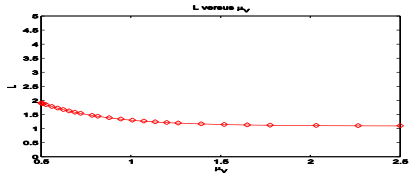


Figure 3: The mean system size versus μ_V with $(\lambda^-, \mu_R) = (0.212, 0.911)$.

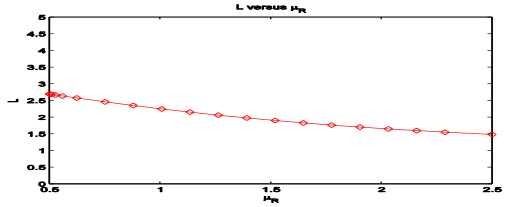


Figure 4: The mean system size versus μ_R with $(\lambda^-, \mu_V) = (0.212, 1.603)$.

So the stationary arrival rates of the positive customers and the negative customers are $\lambda_1 = 0.7250$ and $\lambda_2 = 0.0054\lambda^-$.

In Figures 1 and 2, the mean number of customers in the system is plotted against the parameter λ^- with $\mu_R = 0.911$ and $1/\mu_R = 0$ respectively. We observe that the mean number of customers in the system increases monotonously as the value λ^- increases when $1/\mu_R > 0$, and decreases monotonously as the value λ^- increases when $1/\mu_R = 0$. It is easily explained taking into account the fact that a negative customer not only removes the positive customer being in service but also causes the server breakdown. When the server is reliable, i.e. $1/\mu_R = 0$, the removal of the customer being in service can shorten the queue length. We show in Figures 3 and 4, the influence of the parameters μ_V and μ_R on the mean number of customers L in the system. As is to be expected, L decreases for increasing values μ_V and μ_R.

7 Conclusions

This paper analyzes a MAP/ G / 1 queuing system with negative customer arrival, unreliable server and multiple vacations. By using the supplementary variables method and the censoring technique,we obtain the queue length distributions in steady state. We derive the mean of the busy period based on the renewal theory. Compared to the related work, when threre are no vacations, our results are consistent with the results in [5] and when there are no negative customers, our results are agree with the results in [13]. Hence, our model covers the models considered in [5] and [13]. This queuing system can be applied to the

virtual channel of ATM network Performance analysis is more practical, real and reasonable.

8 Acknowledgement

This work is supported by Provincial Natural Science Foundation of Hunan under Grant 2019JJ50677 and the Program of Hehua Excellent Young Talents of Changsha Normal University and nurturing program of Changsha Normal University.

9 References

[1] Gelenbe E (1991). Product-form queueing networks with negatived and positive customers. Journal of Applied Probability, pp. 656-663. https://doi.org/10.1017/s0021900200042492

[2] Gelenbe E, Glynn P, Sigman K (1991). Queues with negative arrivals. Journal of Applied Probability, pp. 245-250. https://doi.org/10.1017/s0021900200039589

[3] Boucherie RJ, Boxma OJ (1995). The workload in the M/G/1 queue with work removal. Probability in the Engineering and Informational Sciences, pp. 261-277. https://doi.org/10.1017/S0269964800004320

[4] Harrison PG, Patel NM, Pitel E (2000). Reliability modelling using G-queues. European Journal of Operational Research, pp. 273-287.

[5] Li QL, Zhao YQ (2004). A MAP/G/1 Queue with Negative Customers. Queueing Systems, pp. 5-43. https://doi.org/10.1023/B:QUES.0000032798.65858.19

[6] Cao J, Cheng K (1982). Analysis of M/G/1 queueing system with repairable service station. Acta Mathematicae Applicatae Sinica, pp. 113-127. https://doi.org/10.1007/BF01149327

[7] Neuts MF, Lucantoni DM (1979). A Markovian queue with N servers subject to breakdowns and repairs. Managment Science, pp. 849-861. https://doi.org/10.1287/mnsc.25.9.849

[8] Wang J, Cao J, Li Q (2001). Reliability analysis of the retrial queue with server breakdowns and repairs. Queueing Systems, pp. 363-380. https://doi.org/10.1023/a:1010918926884

[9] Harrison PG, Patel NM, Pitel E (2000). Reliability modelling using G-queues, European Journal of Operational Research, pp. 273-287.

[10] Li QL, Ying Y, Zhao YQ (2006). A BMAP/G/1 retrial queue with a server subject to breakdowns and repairs. Annals of Operations Research, pp. 233-270.

[11] Doshi BT (1986). Queueing Systems with Vacations: Survey. Queueing Systems, pp. 29-66.

[12] Sikdar K, Gupta UC (2005). The Queue Length Distributions in the Finite Buffer Bulk-Service MAP/G/1 Queue with Multiple Vacations. Top, pp. 75-103. https://doi.org/10.1007/bf02578989

[13] Kasahara S, Takine T, Hasegawa T (1996). MAP/G/1 Queues Under N-Policy With and Without Vacations, Journal of the Operations Research Society of Japan, pp. 188-212.

[14] Bocharov PP, D'Apice C, Pechinkin AV, Salerno S (2004) Queueing Theory, Boston: Utrecht.

Recurrent Neural Network Training Using ABC Algorithm for Traffic Volume Prediction

Adrian Bosire
Department of Computer Science
Kiriri Womens University of Science and Technology, Kasarani, Kenya
E-mail: bosire.adrian@gmail.com

Student paper

Keywords: deep neural network, recurrent neural network, artificial bee colony

Received: March 8, 2019

This study evaluates the use of the Artificial Bee Colony (ABC) algorithm to optimize the Recurrent Neural Network (RNN) that is used to analyze traffic volume. Related studies have shown that Deep Neural Networks are superseding the Shallow Neural Networks especially in terms of performance. Here we show that using the ABC algorithm in training the Recurrent Neural Network yields better results, compared to several other algorithms that are based on statistical or heuristic techniques that were preferred in earlier studies. The ABC algorithm is an example of swarm intelligence algorithms which are inspired by nature. Therefore, this study evaluates the performance of the RNN trained using the ABC algorithm for the purpose of forecasting. The performance metric used in this study is the Mean Squared Error (MSE) and ultimately, the outcome of the study may be generalized and extended to suit other domains.

Povzetek: Ocena uspešnosti algoritma umetne kolonije čebelje pri optimizaciji ponavljajoče se nevronske mreže.

1 Introduction

The Artificial Bee Colony (ABC) algorithm is based on the intelligent foraging behavior of the honey-bee swarm, which makes it suitable for optimization problems [14]. In his proposal of the ABC algorithm, Karaboga aimed to solve multi-dimensional and multi-modal optimization problems [12]. A function is considered to be multi-modal if it has several local optima. Furthermore, it is multi-dimensional if the local optima are distributed randomly in the search space, essentially complicating the process of finding the optimal solution. The ABC algorithm has been applied to solve many kinds of real-world problems such as leaf-constrained minimum spanning tree problem, flow shop scheduling problem, inverse analysis problem and radial distribution system network reconfiguration problem among others [21], [29].

Basturk and Karaboga [1] evaluated the ABC algorithm based on five multi-dimensional benchmark functions: sphere function, Rosenbrock Valley, Griewank function, Rastrigin function and Step function. The results obtained show that the ABC algorithm is quite robust for multi-modal problems, since it has multi-agents that work independently and in parallel. This is also echoed by the results they obtained after comparing the performance of the ABC with that of the Particle Swarm Optimization algorithm, Particle Swarm Inspired Evolutionary Algorithm and Genetic Algorithm [14].

Karaboga et. al. [17] used the ABC algorithm to train Feed-Forward Artificial Neural Networks with an aim to overcome drawbacks such as getting stuck in local minima and computational complexity. They discovered that the algorithm had good exploration and exploitation capabilities especially in searching for the optimal weight-set which is crucial in training Neural Networks. In this case, exploration refers to the ability to examine the viability of numerous unknown sections in order to discover the global optimum in the search space and exploitation refers to ability to utilize knowledge of the preceding good solutions to find improved solutions.

The data used in this study in the evaluation of the optimized neural network represents the vehicle count at specific junctions of select motorways in the whole of Britain. However, the optimized neural network can be trained for any other road network whose data is available.

The rest of this paper is organized as follows: Section 2 begins with an overview on swarm intelligence followed by Section 3 which explains the fundamental concept of the ABC algorithm. Later, Section 4 looks at the implementation of the ABC algorithm in optimizing the Recurrent Neural Network. In Section 5, we find the experiments and results. Eventually, a summary of the findings of this paper is presented in Section 6.

2 Swarm intelligence

Swarm intelligence refers to the collective intelligence exhibited by the collaborative behavior of social insect colonies or animal societies in pursuit of a defined purpose. This means that the entities that collaborate form a swarm, which is alternatively defined as a set of agents which act on their environment with an aim of solving a distributed problem [23]. These entities work together with a common goal thus increasing their chances of finding the best or optimal solution to the task at hand. In

so doing, they inadvertently enhance the exploration and exploitation of their environment. Furthermore, this process serves to break down the problem into smaller and simpler tasks which are easily solved by sub-groups whose solutions are aggregated to formulate the overall solution. So, the time used to find a solution is decreased exponentially with an increase in the agents involved and also because some of these smaller tasks can be solved concurrently. The dedicated effort of such agents to a single, simplified and well-defined task also minimizes occurrence of errors as may be experienced when a single agent is tasked with the same problem. Therefore, the collective effort is useful in cases where a problem can be compartmentalized into smaller manageable tasks.

Examples of swarm intelligence algorithms include Artificial Bee Colony, Ant Colony Optimization, Particle Swarm Optimization, Immune Algorithm, Bacterial Foraging Optimization, Cat Swarm Optimization, Cuckoo Search Algorithm, Firefly Algorithm, Gravitational Search Algorithm among others [15], [23]. These algorithms are evidence of various assortments of swarms in the world and their varied level of intelligence but self-organization and labor division are key features they collectively possess.

3 Artificial bee colony algorithm

The ABC algorithm is a swarm-based algorithm presented by Karaboga [12]. This algorithm is inspired by the intelligent-search behavior of honeybees, known for their systematic collection of nectar that they process into honey. Nectar (food) is collected from flowers located in the neighboring fields (food sources) away from their hives. The bees communicate with each other by means of a waggle dance so as to share information about the quality of food sources. This information shared among the colony members includes the location and proximity of the food source to the hive, the quality of food source and quantity of food. This majorly governs the foraging range with correct accuracy thus enabling the swarm to direct its efforts to the best food source. Their mutual dependence is pegged on their distinct but partially evolving roles that adapt to the needs of the colony. The needs of the colony, decentralized decision-making and the age of the bees as well as their physical structure serve as a control for their social life. Therefore, self-organization, autonomy, distributed functioning and division of labor constitute the swarms' ability to solve distributed problems as a unit and adapt to any environment. [23], [24], [27].

The intelligence exhibited by the collective behaviour of swarms via local interactions may be characterized into four distinctive features. The firrst one is positive feedback which refers to the creation of convenient structures such as recruitment and reinforcement. Then we have negative feedback that involves counterbalancing of the positive feedback in order to stabilize the collective pattern and avoid saturation The third is fluctuations which involve the variations incurred in form of errors, random task switching among swarm individuals which stimulates creativity and discovery of new structures. Lastly, we have multiple interactions tha refer to the relationship and cooperation between the various agents in the swarm that result in the overall development [17], [18].

The honeybee forage selection model is based on three components: food sources (alternative solutions), employed foragers (active solution seekers) and unemployed foragers (passive solution seekers) made up of onlookers and scouts. In addition, two leading modes of the behavior are expressed: recruitment to a food source and abandonment of a food source. Thus, the position of a food source represents a potential solution to the optimization problem and the quantity of a food source corresponds to the calculated fitness value of the associated solution [12], [13], [14], [26].

In essence, food sources signify the profitability of the proposed solution in terms of complexity involved in attaining it. This complexity is evaluated based on proximity, ease of extraction, energy concentration which is calculated as a probability value. Employed foragers are associated with a particular food source or simply a solution they are working on, whereas, the unemployed foragers are looking for potential food sources to exploit or simply looking out for alternative solutions. Thus, the scouts find alternative food sources while the onlookers establish viable solutions from the information given to them by the employed foragers through the waggle dance.

At the beginning, the number of employed bees and the number of available food sources. Additionally, an employed bee turns into a scout when the position of a food source declines after a predetermined limit of foraging attempts, at that time exploitation ceases. Thus, the employed and onlooker bees usually perform the exploitation whereas the scouts perform the exploration of the search space. This process of foraging can be viewed as a complex problem broken down into many parts and the ultimate task is to find a viable solution since there are many ways in reaching the goal [9], [18], [23]. Let us examine figure 1 as illustrated by Karaboga [12], for a better understanding of this foraging behaviour.

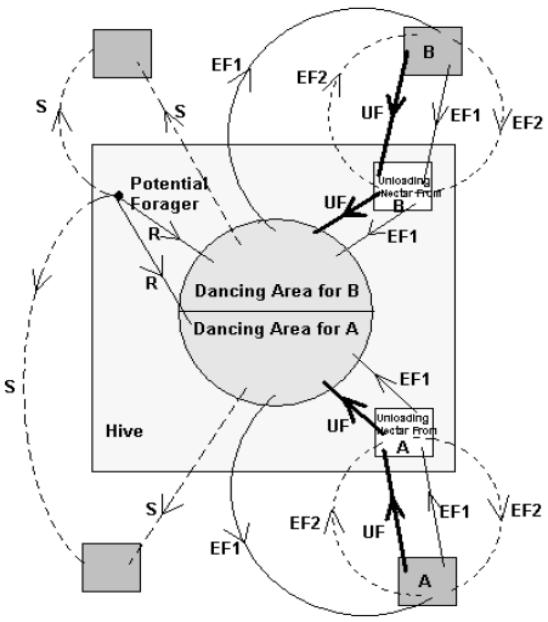


Figure 1: The honeybee nectar foraging behavior [12].

In figure 1 above, there are two discovered food sources: A and B. Any potential forager will always start as an unemployed forager and will not have any knowledge about the food sources around the nest. This limits the prospective options for such a bee to the following:

i. To become a scout and instinctively start searching around the nest for food (S).
ii. To become a recruit after watching the waggle dances for the available food sources (R).

This bee then evaluates the available food sources, memorizes a food source location and immediately starts exploiting it thus becoming an employed forager. The foraging bee takes with it a load of nectar from the source and unloads it to a food store back in the hive after which the bee takes on one of the three roles below:

i. It recruits other bees (onlookers) and returns to the same food source (EF1).
ii. It continues to forage at the same food source without recruiting other bees (EF2).
iii. It becomes an uncommitted follower after abandoning the food source (UF).

Therefore, this formulates the procedure of the ABC algorithm which is separated into five distinct phases; Initialization phase, Employed bee phase, Probabilistic selection phase, Onlooker bee phase and the Scout bee phase [12], [23]:

i. Initialization Phase

The Food Source locations are randomly initialized within the search space as calculated using equation (1) below.

$$x_{ij} = x_j^{min} + rand(0,1)(x_j^{max} - x_j^{min}) \quad (1)$$

where $i = 1, 2, \ldots, SN$ and SN indicates the number of Food Sources (equal to half of the bee colony);
$j = 1, 2, \ldots, D$ and D is the dimension of the problem;
x_{ij} represents the parameter for i^{th} employed bee on j^{th} dimension, meaning that they are dependent on each other;
x_j^{max} and x_j^{min} are upper and lower bounds of x_{ij}.

ii. Employed Bee Phase

Every Employee Bee is assigned to the resultant Food Source generated by equation (2) below for further exploitation.

$$v_{ij} = x_{ij} + \varphi_{ij}(x_{ij} - x_{kj}) \quad (2)$$

where k is a neighbor of i, $i \neq k$;
φ_{ij} is a random number in the range $[-1, 1]$ to control the production of neighbor solutions around x_{ij};
v_{ij} is the new solution for x_{ij}.

The value of the new Food Source is measured using a fitness value calculated by equation (3) below.

$$fit_i = \begin{cases} \frac{1}{1 + absf_i}, & f_i \geq 0 \\ 1 + abs(f_i), & f_i < 0 \end{cases} \quad (3)$$

where $absf_i$ is the absolute objective function associated with each Food Source;
fit_i is the fitness value.

The two food sources x_{ij} (Original Food Source) and v_{ij} (New Food Source) are compared and the best is chosen based on a greedy selection of their fitness values.

iii. Probabilistic Selection Phase

Then, a probability value for each Food Source is calculated using equation (4) which is useful for Onlooker Bees when they evaluate the viability of a Food Source amongst the available options.

$$p_i = \frac{fit_i}{\sum_{j=1}^{N} fit_j} \quad (4)$$

where fit_i is the fitness value of i-th solution;
p_i is the selection probability of $\boldsymbol{i}$-th solution.

iv. Onlooker Bee Phase

The Employed Bees advertise the viability of their Food Sources to the Onlooker Bees which select a Food Source to exploit based on the fitness and probability values associated with it i.e., the more fitness, the higher the probability. The Food Sources that are picked are further exploited using equation (2). This improves the solution and their fitness values are also calculated using equation (3). Once again, to yield an improved solution, a greedy selection process is performed on the original and new Food Sources, similar to Employed Bee Phase.

v. Scout Bee Phase

The Employed Bee for a Food source that doesn't generate better results over time becomes a Scout Bee and the Food Source is abandoned. This leads to the random generation of a new Food Source in the search space using equation (1). Subsequently, the Employed bee phase, Probabilistic selection phase, Onlooker bee phase and Scout bee phases will execute until termination criterion is satisfied. The best food source solution is obtained as output. Note that the steps of the algorithm presented in section 4 are quite elaborate than the fore mentioned summary. [12], [13], [15], [18].

4 RNN training using ABC algorithm

Artificial Neural Networks are based on the simulated network of biological neurons in which neurons are the essential computational units [22]. Hence, the underlying concept is to train a mathematical model so that it can reproduce some physical phenomena or make some predictions. The model is presented with training samples that are the actual outputs of the studied system corresponding to the actual inputs of the problem. Later, the error obtained between the actual and the predicted value serves as the metric for measuring the performance of the algorithm in terms of prediction [5].

Artificial Neural Networks can broadly be categorized into Shallow Neural Network and Deep Neural Network techniques. Shallow Neural Networks generally have only one hidden layer as opposed to Deep Neural Networks which have several levels of hidden layers. Therefore, Deep Neural Networks utilize functions whose complexity is of a higher magnitude contrary to

Shallow Neural Networks, given that all resources remain constant [3].

Shallow Neural Network (SNN) techniques contain less than two layers of nonlinear feature transformations. Examples of the SNN techniques are Conditional Random Fields (CRFs), Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs), Maximum Entropy (MaxEnt) models, Logistic Regression, Kernel Regression, Multi-Layer Perceptron's (MLPs) with a single hidden layer including Extreme Learning Machines (ELMs). SNN techniques effectively solve well-constrained problems due to their limited modeling and representational power which poses a challenge when dealing with complicated real-world applications. A well-constrained problem is one for which a function is to be minimized or maximized with respect to well defined constraints [3], [6].

Deep Neural Networks (DNN) are Artificial Neural Networks composed of several interconnected hidden layers. These hidden layers have multiple hidden perceptrons between the network input layer and its network output for computational use. Dynamic environments require Deep Neural Network techniques which are useful in extracting complex structure and building internal representation. Examples of DNNs are Recurrent Neural Network (RNN), Convolutional Neural Networks (Conv.Net), Deep Boltzmann Machines (DBM), Deep Belief Networks (DBN) [30].

So, the basic concept behind Artificial Neural Networks owes to their imitation of biological neurons as shown in figure 2 which is an elementary neuron with several inputs and one output. Here, each input *x* is fed to the next layer, in our case an output layer *y*, with an appropriate weight *w*. The sum of the weighted inputs and the bias forms the input to the transfer function *f*. The bias is a threshold that represents the minimum level that a neuron needs for activating and is represented by *b*. Neurons can use any differentiable transfer function *f* to generate their output. Therefore, in multi-layer networks, the input values to the inputs of the first layer, allow the signals to propagate through the network, and read the output values where output of the i^{th} node can be described by the function in Eq. 4.1 below [25], [28].

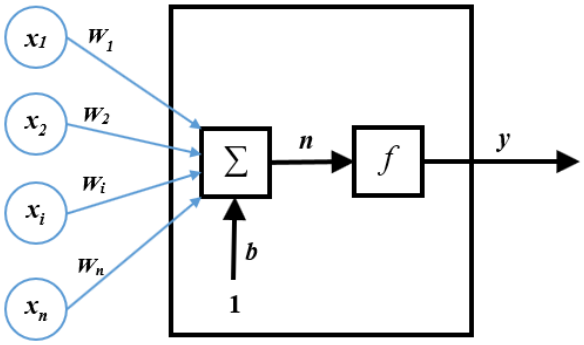


Figure 2: Representation of an Elementary Neuron.

$$y_i = f_i(\sum_{j=1}^{n} w_{ij}x_j + b_i) \quad (5)$$

where y_i is the output of the node;
x_j is the j^{th} input to the node;
w_{ij} is the connection weight between the node and input x_j;
b_i is the threshold (or bias) of the node;
f_i is the node transfer function.

Multilayer networks often use the sigmoid transfer function which generates outputs between 0 and 1 as the neuron's net input goes from negative to positive infinity. This is used for models where we have to predict the probability as an output. Hence, its suitability because the probability of real-world entities exist in the range of 0 and 1. Sigmoid output neurons are often used for pattern recognition, clustering and prediction problems.

The information from a layer to the next one is transmitted by means of the activation function, represented in equation (6). The activation function relies on the weighted sum and bias to make a calculation on whether a neuron will be activated or not, thus introducing non-linearity to the network. This non-linear transformation performed on the inputs and sent through the network enables it to learn and perform complex tasks.

$$y = f(n) = \frac{1}{1+e^{-n}} \quad (6)$$

The main goal is to minimize the cost function by optimizing the network weights. The fundamental idea of this optimization approach is to individually interpret and change the weight values. Also, note that dynamic environments present a relatively higher network complexity which suggests the need for Deep Neural Networks. Therefore, the data presented to the network has to be split into three sets; training set, validation set and the testing set. This facilitates the training, verification and evaluation of the networks' performance. Furthermore, the complexity of the challenge is represented by the Mean Squared Error (MSE) in equation (7). The MSE is obtained while comparing the target input against the predicted output could determine the number of hidden layers. The optimization of the network is achieved by minimizing the MSE which is essentially a network error function. Henceforth, the training algorithm is used to find the optimal weights that are used for initializing the Neural Network. In this case, the ABC algorithm is used to find the precise weights that enable the network connections to make accurate decisions. The algorithm uses a cost function as a measure for our progress in determining the right weights [19], [25].

$$E\big(w(t)\big) = \frac{1}{n}\sum_{k=1}^{n}(d_k - O_k)^2 \quad (7)$$

where, $E\big(w(t)\big)$ is the error at the t^{th} iteration;
$\big(w(t)\big)$, the weights in the connections at the t^{th} iteration;
d_k and O_k represent the desired and the actual values of k^{th} output node;
k is the number of output nodes;
n is the number of inputs.

A Recurrent Neural Network (RNN) is an extension of the conventional feed-forward neural network described above. The major difference is that RNNs have cyclic connections which make them reliable for modeling time-series data in dynamic environments. This means

that at any given the output is related to the present input and the input at previous timestamps. Therefore, we build on the concept above of the elementary neuron in relation to the RNN. Here, we have the input sequence denoted by $x = (x_1, x_2, ..., x_t)$, the hidden layer denoted by $h = (h_1, h_2, ..., h_t)$ and the output vector sequence denoted by $y = (y_1, y_2, ..., y_t)$. Usually the RNN calculates the hidden vector sequence h using equation (8) and the output vector sequence y using equation (9) with t = 1 to T [20];

$$h_t = f_t\,(w_{xh}x_t + w_{hh}h_{t-1} + b_h) \quad (8)$$

$$y_t = f_t\,(w_{hy}h_t + b_y) \quad (9)$$

where function f_t is the activation function;
w is a weight matrix;
b is the bias term.

However, the Long Short-Term Memory (LSTM) architecture is preferable because it resolves the underlying vanishing and exploding gradient problems of the traditional RNN. The LSTM – RNN uses three gates that form a cell which consequently solves the problems mentioned above thus making the network robust. Thus, the LSTM cell replaces the recurrent hidden cell in Eq. 4.4 above. The equations to compute the values for the three gates are described below [11], [20].

$$i_t = f_t\,(w_{xi}x_t + w_{hi}h_{t-1} + w_{ci}c_{t-1} + b_i) \quad (10)$$

$$g_t = f_t\,(w_{xg}x_t + w_{hg}h_{t-1} + w_{cg}c_{t-1} + b_g) \quad (11)$$

$$c_t = f_t c_{t-1} + i_c \tan h\,(w_{xc}x_t + w_{hc}h_{t-1} + b_c) \quad (12)$$

$$O_t = f_t\,(w_{xo}x_t + w_{ho}h_{t-1} + w_{co}c_{t-1} + b_o) \quad (13)$$

$$h_t = O_t \tan h\,(c_t) \quad (14)$$

Where, f_t is the logistic sigmoid function;
i, g, o and c are respectively the input gate, forget gate, output gate and cell state;
w_{ci}, w_{cg} and w_{co} are denoted weight matrices for peephole connections.

In LSTM – RNN, the input gate ***i,*** the forget gate ***g***, and the output gate ***o*** control the information flow. The input gate decides the ratio of input which has an effect when calculating the cell state, c. The forget gate calculates the ratio of the previous memory h_{t-1} using equation (11) and decides whether to pass it onwards or not. The result obtained is used for determining the cell state in equation (12). The output gate which is based on equation (13) determines whether pass out the output of the memory cell or not. This process as represented by the ratios from the three gates is denoted by equation (14) and also depicted diagrammatically in the figure 3 [20].

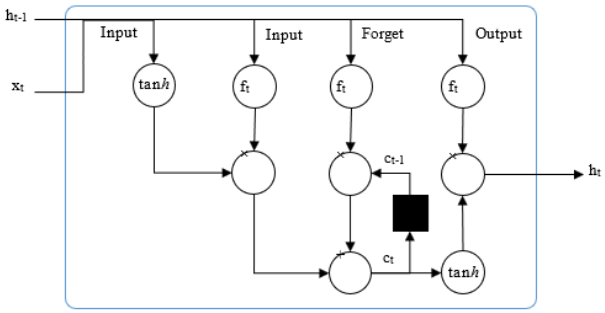


Figure 3: Long-Short Term memory Cell.

Therefore, the algorithm below outlines the optimization process for the deep neural network using the ABC algorithm [10], [12], [19], [25].

1. Set Cycle=0.
2. Load training samples from dataset.
3. Initialize a population of scout bee with random solution x_i, $i = 1,2, ..., SN$ using equation (1).
4. Evaluate fitness (fit_i) of the population using equation (3)

a. Initialize weight and bias for the Recurrent Neural Network

5. Set Cycle=1: while Maximum cycle not reached, repeat step 6 – step 12
6. FOR each employed bee {
   Produce new solution v_i by using equation (2)
   Calculate the value fit_i on the new population
   Apply greedy selection process between x_{ij} and v_{ij}}
7. Calculate the probability values p_i for the solutions (x_i) using equation (4)
8. FOR each onlooker bee {
   Select a solution x_i depending on p_i
   Produce new solution v_i
   Calculate the value fit_i
   Apply greedy selection process}
9. If there is an abandoned solution for the scout then replace it with a new solution which will be randomly produced by equation (1)
10. Memorize the best solution so far
11. Update new weight and bias for the Recurrent Neural Network
12. Increment Cycle + 1 until Cycle=MCN

where x_i represents a solution;
fit_i is the fitness value of x_i;
v_i indicates a neighbor solution of x_i;
p_i is the probability value of x_i;
MCN is the maximum cycle number in the algorithm.

Remember that at the beginning, one half of the colony consists of onlooker bees and the second half constitutes the employed bees which are equal to the number of food sources (viable solutions) and any employed bee whose food source has been exhausted becomes a scout bee. Therefore, the algorithm starts by generating a randomly distributed initial population (SN food source positions), where SN denotes the size of population. Each solution xi (i = 1, 2, ..., SN) is a D-dimensional vector. D being the number of optimization parameters. After initialization, the population of the solutions is subjected to repeated cycles, C = 1, 2, ..., MCN, of the search process until a termination criterion is achieved. Each cycle of the search consists of three steps: engaging the employed bees with their food sources and evaluating their viability; sharing the food sources viability information with the onlookers which select a food source and again assess its viability; determining the scout bees and sending them out randomly to explore new food sources. An employed bee produces a modification on the solution in its memory depending on the probability and fitness tests. Thereby, generating optimal weights that serve to minimize the cost function and with each cycle

the RNN is adequately trained with varying parameters using the ABC algorithm until optimal conditions are met [10], [18], [19].

5 Experiments and results

During the training phase, the Recurrent Neural Network is presented with a set of the training data from the dataset and the input weights are adjusted by using the ABC algorithm as a learning algorithm. The dataset can be acquired from the Road Traffic Statistics website for Great Britain [7]. The purpose of the weight adjustment is to enable the RNN to learn so that it would adapt to the given training data [10]. The dataset also has to be split to a suitable ratio to enable the training of the network, validation and testing of the results obtained. Thereafter, the performance of the network is evaluated based on the Mean Squared Error (MSE) obtained between the desired output and the actual output thus testing the validity of the network in terms of its prediction efficiency. Figure 4 below depicts the performance graph obtained on execution of the algorithm in MATLAB [2].

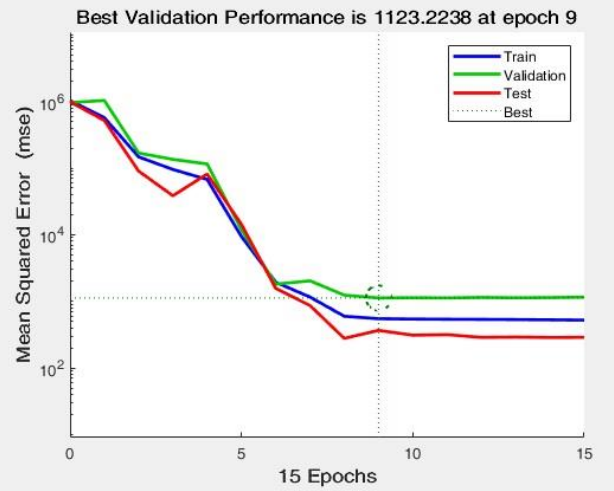


Figure 4: Performance of the ABC Optimized RNN.

The figure 4 represents the best validation performance of the network. On several runs of the algorithm the MSE obtained was 1.1232e3. This is the value obtained on epoch 9 after which the error gradually starts to increase due to overfitting but in this case, it gradually maintains a constant level. In other experiments, the MSE of the RNN before it was optimized was 3.853e3 [4]. The difference between the two MSEs basically shows that the ABC algorithm is actually efficient in terms of optimization. Generally, lower MSEs translate to high accuracy. Graphically, this is seen in the regression plots for the dataset in figure 6.

Figure 5 shows the respective regression values of the three different sets of the dataset. Splitting of the dataset helps with the early stopping of the network in order to achieve its generalization capability. The three sets of data all obtain value greater than 0.9 and the aggregate regression value is 0.93625 which borders 1. This shows a high relationship between the desired outputs and the obtained outputs, which shows a high accuracy in the

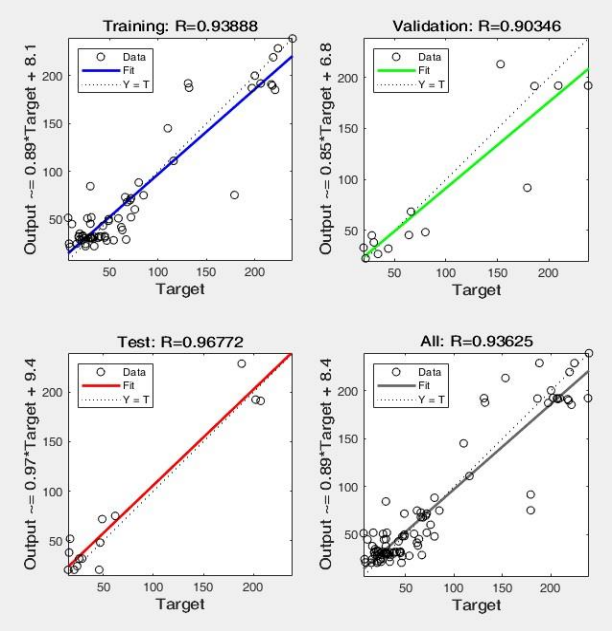


Figure 5: ABC Optimized RNN regression plot.

networks ability to forecast efficiently. Furthermore, there is a high cross-correlation between the input data and the error time-series as depicted in the graph in figure 6 below.

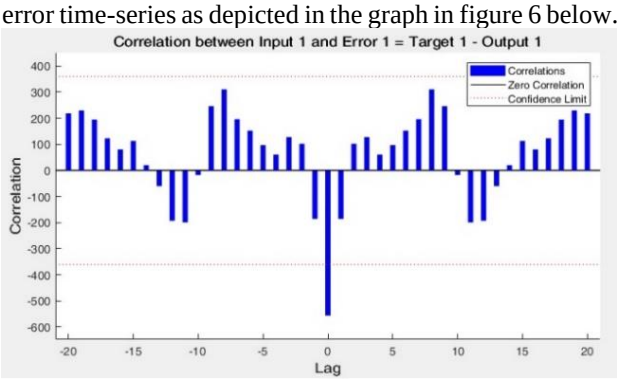


Figure 6: Correlation between the input and the output error.

The figure 6 above means that the network is able to model the predictive characteristics of the time-series lag which is the difference between the expected and the actual values. This correlation is depicted in the figure and the values fall in between the acceptable confidence limits as shown by the dotted red line. This is further exemplified in the time series plot of figure 7 below which shows the relationship between the predicted values and the actual values

The figure 7 above shows the desired output values plotted against the actual values obtained by the RNN optimized by the ABC algorithm. This time-series graph shows the level of accuracy that can be obtained during prediction with a well-trained RNN. The high efficacy of the ABC algorithm is also depicted in the graph regardless of one incorrectly predicted value. However, the other values fall between the confidence limits and as such with further training and fine-tuning of the parameters the RNN can actually produce reliable results. This means that the generalized model can actually produce accurate predictions. The values of the RNN after optimization

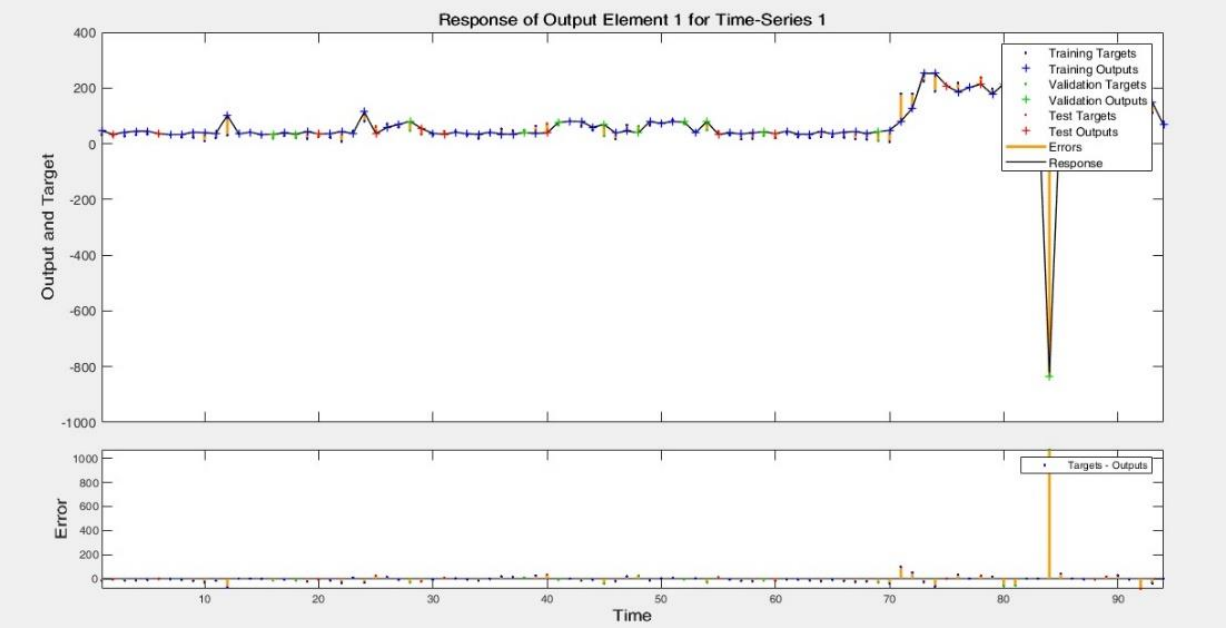


Figure 7: Time-series response of the RNN output values.

Training Algorithm	Hidden Layer Size	Performance (MSE)		
		Training	Validation	Testing
Artificial Bee Colony (ABC) Algorithm	20	721.2039	998.2733	1689.7351
	40	412.1763	569.0172	673.4512
	60	359.8409	920.2444	1.0031e+01
	80	492.2976	1.1457e+01	2.4915e+01
	100	540.0012	2.4345e+01	2.7031e+02

Table 1: MSEs obtained by the ABC optimized Recurrent Neural Network.

using the ABC algorithm are illustrated in the table 1 below.

The results in table 1 above reflect the MSEs obtained during the training, validation and testing phases of the experiment. The optimal MSE for the training phase was 359.8409, the validation phase had an optimal MSE of 569.0172 and the optimal MSE at the Test phase was 673.4512. These values show that the error rate reduced gradually with an increase in the number of hidden layers. Other experiments have been performed using other algorithms for optimization of the deep neural networks. These algorithms include the Levenberg-Marquardt Backpropagation algorithm which had an optimal MSE of 360.2578 at the training phase, the Scaled Conjugate Gradient Backpropagation algorithm which had a least MSE of 480.9656 at the validation phase and the Resilient Backpropagation algorithm which had 467.9015 as the MSE at the testing phase [4]. The ABC trained RNN has peak performance when the hidden layer size is between 40 and 80 given an input vector size of 500. In comparison to the fore-mentioned training algorithms, the ABC algorithm surpasses the other training algorithms in similar conditions.

6 Conclusion

It is evident from the results that the ABC algorithm out-performs the backpropagation algorithms. However, the parameter settings for the algorithm need to be refined for the model to be generalized. Moreover, different architectures of other deep neural networks can be implemented especially in distributed computing environments so as to sustain a greater number of the hidden layers or even produce a sustainable hybrid thereof. Furthermore, deep neural networks need to be optimized so as to enhance the practicability of a model that yields reliable forecasting in dynamic environments.

7 Acknowledgement

This work was supported by Kiriri Womens University of Science and Technology.

References

[1] Basturk, B., & Karaboga, D. (2006). An Artificial Bee Colony Algorithm (ABC) for Numeric Function Optimization. *IEEE Swarm Intelligence Symposium.* Indianapolis, Indiana, USA.

[2] Beale, M. H., Hagan, M. T., & Demuth, H. B. (2018, February). MATLAB R2018a. *Neural Network Toolbox Version 11.1*. MathWorks Inc. Retrieved from: https://www.mathworks.com/products/deep-learning.html

[3] Bianchini, M., & Scarselli, F. (2014, August). On the complexity of Neural Network Classifiers: A Comparison Between Shallow and Deep Architectures. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(8), 1553-1565.

https://doi.org/10.1109/TNNLS.2013.2293637

[4] Bosire, A., Okeyo, G., & Cheruiyot, W. (2018, October). Performance of Deep Neural Networks in the Analysis of Vehicle Traffic Volume. *International Journal of Research and Scientific Innovation, 5*(10), 57-66.

[5] De Luca, G., & Gallo, M. (2017). Artificial Neural Networks for forecasting user flows in transportation networks: literature review, limits, potentialities and open challenges. *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems* (pp. 919-923). Naples, Italy: IEEE. https://doi.org/10.1109/MTITS.2017.8005644

[6] Deng, L., & Yu, D. (2013). Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing, 7*, 197-387. https://doi.org/10.1561/2000000039

[7] Department for Transport. (2018, June). *Traffic counts*. Retrieved June 2018, from Traffic counts: https://www.dft.gov.uk/traffic-counts/about.php

[8] Garro, B. A., & Vázquez, R. A. (2015). Designing Artificial Neural Networks Using Particle Swarm Optimization Algorithms. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2015/369298

[9] Haris, P., Gopinathan, E., & Ali, C. (2012, July). Artificial Bee Colony and Tabu Search Enhanced TTCM Assisted MMSE Multi-User Detectors for Rank Deficient SDMA-OFDM System. *Wireless Personal Communications, 65*(2), 425-442. https://doi.org/10.1007/s11277-011-0264-0

[10] Hassim, Y. M., & Ghazali, R. (2012). Training a Functional Link Neural Network Using an Artificial Bee Colony for Solving a Classification Problems. *Journal of Computing, 4*(9), 110-115.

[11] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[12] Karaboga, D. (2005). *An Idea Based on Honey Bee Swarm for Numerical Optimization.* Technical Report, Erciyes University, Computer Engineering Department, Turkey.

[13] Karaboga, D., & Akay, B. (2009, August). A comparative study of Artificial Bee Colony algorithm. *Applied Mathematics and Computation, 214*(1), 108-132. https://doi.org/10.1016/j.amc.2009.03.090

[14] Karaboga, D., & Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: Artificial Bee Colony (ABC) algorithm. https://doi.org/10.1007/s10898-007-9149-x

[15] Karaboga, D., & Basturk, B. (2008). On the performance of artificial bee colony (ABC) algorithm. *Applied Soft Computing*, 687–697. https://doi.org/10.1016/j.asoc.2007.05.007

[16] Karaboga, D., & Ozturk, C. (2009). Neural Networks training by Artificial Bee Colony algorithm on pattern classification. *Neural Network World, 19*(3), 279–292.

[17] Karaboga, D., Basturk, B., & Ozturk, C. (2007). Artificial Bee Colony (ABC) Optimization Algorithm for Training Feed-Forward Neural Networks. In *Modeling Decisions for Artificial Intelligence* (pp. 318-329). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73729-2_30

[18] Karaboga, D., Gorkemli, B., Ozturk, C., & Karaboga, N. (2012). A comprehensive survey: Artificial Bee Colony (ABC) algorithm and applications. *Artif Intell Rev, 42*, 21–57. https://doi.org/10.1007/s10462-012-9328-0

[19] Kayabasi, A. (2018). An Application of ANN Trained by ABC Algorithm for Classification of Wheat Grains. *International Journal of Intelligent Systems and Applications in Engineering, 6*(1), 85-91. https://doi.org/10.18201/ijisae.2018637936

[20] Kim, J., Kim, J., Thu, H. L., & Kim, H. (2016). Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection. 1-5. https://doi.org/10.1109/PlatCon.2016.7456805

[21] Koc, E., Ersoy, N., Andac, A., Camlidere, Z. S., Cereci, I., & Kilic, H. (2012). An empirical study about search-based refactoring using alternative multiple and population-based search techniques. In E. Gelenbe, R. Lent, & G. Sakellari (Ed.), *Computer and information sciences II* (pp. 59-66). Springer, London. https://doi.org/10.1007/978-1-4471-2155-8_7

[22] Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science, 1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

[23] Kumar, A., Kumar, D., & Jarial, S. K. (2017). A Review on Artificial Bee Colony Algorithms and Their Applications to Data Clustering. *Cybernetics and Information Technologies, 17*(3). https://doi.org/10.1515/cait-2017-0027

[24] Omkar, S., & Senthilnath, J. (2009). Artificial Bee Colony for Classification of Acoustic Emission Signal Source. *International Journal of Aerospace Innovations, 1*(3), 129-143. https://doi.org/10.1260/175722509789685865

[25] Ozturk, C., & Karaboga, D. (2011). Hybrid Artificial Bee Colony Algorithm for Neural Network Training. *2011 IEEE Congress of Evolutionary Computation (CEC)*, 84-88. https://doi.org/10.1109/CEC.2011.5949602

[26] Seyyed, R. K., Maleki, I., Hojjatkhah, S., & Bagherinia, A. (2013, August). Evaluation of the Efficiency of Artificial Bee Colony and Firefly Algorithm in Solving the Continuous Optimization Problem. *International Journal on Computational Sciences & Applications, 3*(4).

[27] Shukran, M. A., Chung, Y. Y., Yeh, W.-C., Wahid, N., & Zaidi, A. M. (2011, August). Artificial bee colony based data mining algorithms for classification tasks. *Modern Applied Science, 5*(4), 217–231. https://doi.org/10.5539/mas.v5n4p217

[28] Vazquez, R. A., & Garro, B. A. (2015). Training Spiking Neural Models Using Artificial Bee Colony. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2015/947098

[29] Xiangyu, K., Liu, S., & Wang, Z. (2013). An Improved Artificial Bee Colony Algorithm and Its Application. *International Journal of Signal Processing, Image Processing and Pattern Recognition, 6*(6), 259-274. https://doi.org/10.14257/ijsip.2013.6.6.24

[30] Yann, L., & Ranzato, M. (2013). Deep learning tutorial. Tutorials in International Conference on Machine Learning (ICML'13).

List of Abbreviations

ABC	Artificial Bee Colony
Conv.Net	Convolutional Neural Network
CRF	Conditional Random Field
DBM	Deep Boltzmann Machine
DBN	Deep Belief Network
DNN	Deep Neural Network
EF	Employed Forager
ELM	Extreme Learning Machine
GMM	Gaussian Mixture Model
LSTM	Long-Short Term Memory
MaxEnt	Maximum Entropy
MCN	Maximum Cycle Number
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
NN	Neural Network
R	Recruited Bee
RNN	Recurrent Neural Network
S	Scout Bee
SNN	Shallow Neural Network
SVM	Support Vector Machine
UF	Unemployed Forager

Study of Computerized Segmentation & Classification Techniques: An Application to Histopathological Imagery

Pranshu Saxena
IK Gujral Punjab Technical University, Kapurthala, India
E-mail: pranshusaxena@gmail.com

Anjali Goyal
Guru Nanak Institute of Management and Technology, Ludhiana, India
E-mail: anjali.garg73@gmail.com

Student paper

Keywords: computer-assisted diagnosis system, classification, digital pathology, histopathological images, images analysis, segmentation

Received: January 9, 2018

Recent trends with histopathological imagery led to rapid progress towards quantifying the perceptive issues, while prognostic, due to subjective variability among readers. This variability leads to distinguished prognosis reports and generates variability in treatment as well. Latest advancements in image analysis tools have allowed the powerful computer-assisted diagnostic system to assist oncologist in their diagnosis process on radiological data. The main goal of this study is to understand and address the challenges associated with the development of image analysis techniques for computer-aided interpretation of histopathology imagery. We are analyzing indicative characteristics like texture heterogeneity and morphological characteristic on a various scale of lymphomas like Follicular, Neuroblastoma, Breast, and Prostate tissue images for classifying them into respective grades. The study shows a systematic survey of the computational steps, which includes a recent scenario of diagnosis process to classify these lymphomas into respective grades along with its limitations, followed by it shows the pre-requisite of the computer-assisted diagnosis system and finally explains various segmentation techniques based on image descriptor and subsequent classification of biopsy into respective grades. This paper reviews recent state of the art technology for histopathology and briefly describes the recent development in histology and its application towards quantifying the perceptive issue in the domain of histopathology being pursued in the United State and India.

Povzetek: Študij tehnik računalniške segmentacije in klasifikacije: uporaba na histopatoloških posnetkih.

1 Introduction

In the current scenario, the prognosis of lymphomas is done manually by visual analysis of the tissue samples; obtained from the biopsy of patients. In clinical medicine, a biopsy sample is obtained from a suspicious histological section, placed onto glass slides to be examined under the microscope. Later, biopsy undergoes with staining process preferable Hematoxylin and Eosin (H&E Stained). H&E stain is the microscopic study of biological tissues, which will become the gold standard in the prognosis of considerable number of pathologies and for the identification of therapeutic effects. Combination of H&E stain produces blue, violet and red colors. It provides information about tissues and cells with a high level of detail [1]. Another type of staining known as Immunohistochemical (IHC) staining, is widely used in the diagnosis of abnormal cells and is used to diagnose and track specific cellular anomalies, such as cancers, by identifying those proteins that are specifically found in affected cells. From H&E Stained images, pathologists inspect morphological characteristic (based on staining) that are indicative of the presence of cancer structures at various scales and determine how closely these structures resemble those in healthy vs. diseased tissues. Sooner the presence of cancer is confirmed, the grading process starts.

The morphological characteristic and texture inhomogeneity of these tissue cells is highly relevant and critically required to predict which patients may be inclined to disease and predicting disease outcome and chances of early survival. This study includes prostate (CaP), breast (BC), neuroblastoma (NB) and follicular lymphoma (FL) tissues histology for critically reviewing the recent state of the art for computer-assisted diagnosis technology, and analyzes the modern state of progress, application of novel image analysis technology and highlights various histopathology related issues.

1.1 Need for quantitative image analysis for disease grading

In the current scenario, pathologists play a vital role in examining the digital histopathological image. This examining is done under the microscope but analyzing all

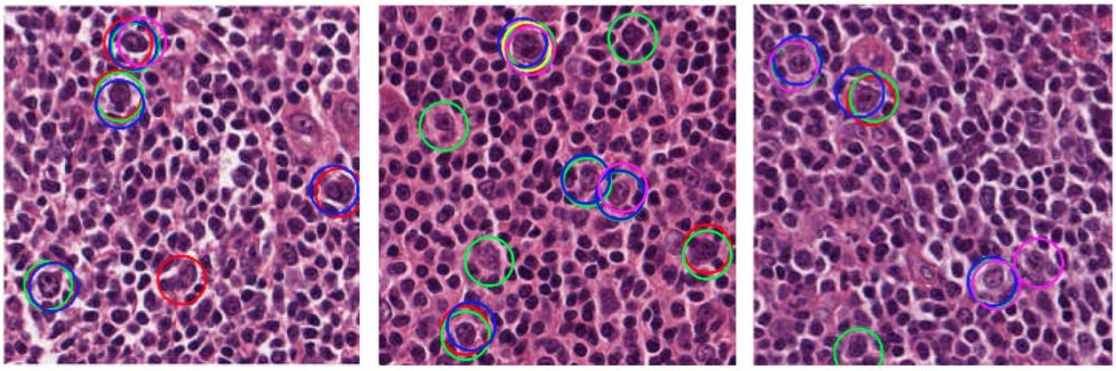


Figure 1: Sample Area of interest taking from 512×512 pixels H&E stained FL tissues images; five different oncologists indicating by circles of different colors identify CB cells. [Olcay Sertel et al. 2010] [3].

these microscopic digital images by the pathologist is a very tedious process, and sometimes leads to incorrect conclusions due to several reasons;

- It is not practical to examine every region of the tissue slide under the microscope at high magnifications (e.g., 40×). For cell diagnosis, due to high CB/HPF ratio (CB-Centro blast & HPF- High Power Field i.e. 0.159mm^2) in case of FL histology [2].
- The resulting diagnosis can vary considerably among different readers i.e. subjectivity issue.
- Biological variation, heterogeneous intensity, uneven staining, illumination, multiple regions of interest and overlapping the cell nuclei have made prognosis procedure a major impediment.
- There is also an essential requirement for the quantitative-based grading system to ease the workload on pathologists/oncologists by detailed study for non-harmful regions (including all type of cancers); As per a study [1], approximately 80% of 10 lakh cancer biopsies done every year which results in a negative response. This implies that there was, no need to do the biopsy at all.

Moreover, quantitative analysis can be very useful for research application like drug discovery, biological mechanisms in disease, identifying the pattern of genetic abnormalities in cancerous nuclei. But recent research says, [39] analyzing transformation in biological tissues remains a challenge due to unavailability of the robust biomarker in routine clinical practice. Indeed, several genetic alterations, IHC stained or H&E stained markers have been reported in previous years associated with biological transformation but haven't been sufficiently validated to warrant their assessment outside of the research setting. Use of computer-assisted diagnosis in pathology can substantially enhance the efficiency and accuracy of pathologists in decision-making, which indeed favors the patients [1]. It is emphasized that the automated computer-assisted diagnosis system should never be considered a replacement for oncologists. Instead, it should only be used as assistance to the decision-making mechanism. In case of a disagreement between the computerized system and the oncologists rating, the final decision is that of the human doctors i.e. oncologists.

With the recent trends and advancement in the CAD system, this review discusses and depict different methods suggested in the literature for segmentation and subsequent classification of FL, BC, NB, & CaP histopathological images. The main emphasis of identifying the most often methods of lymphoma images segmentation, such as thresholding, fourier based transformation, region-based segmentation, k-means clustering, statistical shape model, texture-based segmentation, and methods for their classification into respective grades, such as supervised and unsupervised clustering, laplacian eigen-map classifiers, k-nearest neighbor, rule-based classifiers and neuro-fuzzy inference system. The major contribution of this study is found in the discussion of the main processing techniques of lymphoma histological images, therefore, providing directions for future research.

1.2 Organization of this paper

We have organized this paper to follow the general image analysis procedure for histopathological imagery. These analysis procedures are usually applicable to all histopathological imagery. In section 2, we present the definition and details about quantitative criteria for disease grading on the various scale for FL, NB, CaP, and BC. In section 3, we have reviewed various histological slide preparation process followed by image pre-processing steps such as colour normalization, image representation. Later, a systematic review is placed for segmentation and classification for various lymphomas followed by inference from the review of literature in the table. Discussion and future directions for research of lymphomas image segmentation considering limitation for published articles, followed by the conclusive remark is presented in Section 4.

2 Quantitative criteria for disease grading

In the current scenario prognosis of the diseases such as CaP, BC, FL & NB is done manually by visual analysis of the tissue samples, obtained from the biopsy of patients. This visual grading is very tedious and sometimes leads to under and over treatment due to inter-reader and intra-reader variability; this result indicates incapable situation for the patient, figure 1 shows distinguished prognosis report of the same biopsy of FL sample. In Figure 1, the

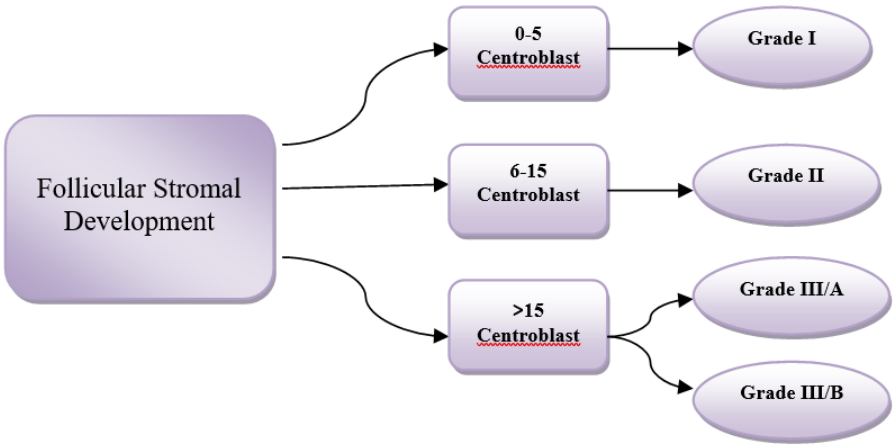


Figure 2: WHO classification of FL based on centroblast/HPF; Grade I & Grade II are considered as low-risk categories, while grade III belongs to high categories.

prognosis process is explained using three different follicular tissue samples by 5 distinguished oncologists. The oncologists were asked to locate the area of interest. All five oncologists have marked their perceptive views with different colors on the tissue sample, which clearly shows a different perceptive view with different oncologists.

A computer-assisted diagnosis system for digitized histopathology is pre-requisite that is employed for disease prognostics, allowing the oncologist to predict which patients may be susceptible to disease and predicting disease outcome and chances of survival. Automated image segmentation of cell nuclei and subsequent classification of histopathological images is widely research topic. Several distinguished researchers have been developing the algorithms for the automated segmentation and classification. There have been few attempts to automatically detect the grading and to newer methods continue to be investigated. The need for quantitative image analysis in the context of some specific diseases (like prostate, breast, follicular lymphoma & neuroblastoma) is described in the next section.

2.1 Grading criteria for follicular lymphoma by WHO

The World Health Organization (WHO) recommends Histopathological grading of Follicular lymphoma based on the number of large malignant cells, namely CB, per standard 40x magnifications High Power Field (HPF) of 0.159 〖 mm〗 ^2. In this method, CB's are manually counted and the average of CB/HPF is reported [2, 4].

According to WHO, high power field of H&E stained tissue sections (under the microscope) classify the biopsy into one of the three histopathological grades according to the average CB count per HPF, this grading criterion is plotted in the figure 2. In grades, I and II, the proportion of small cell (centrocytes) is predominant, whereas grade III features a greater proportion of large cells (centroblast). The histopathological examination guides the oncologists in making timely decisions and on the required therapy [2, 4-22].This clinical relevance of grading system is still being debatable [38]; Grade I, II, IIIA are sluggish and incurable, while Grade IIIB belongs to most aggressive but curable. In the above classification, Grades I and II belong to quantitative issue (count the number of CB/HPF) while classification of Grade III in its respective subgroups requires qualitative issue as CB/HPF ratio remains same (>15 in both cases Grade IIIA &B). The only subjective issue which distinguishes between Grade IIIA & B is that grade IIIA shows the presence of centrocytes, whereas in IIIB the follicles consist almost entirely of the centroblastic cell. Although patients suffering from indolent FL (Grade I, II & IIIA) typically live for many decades with minimal or no treatment, while patients with aggressive FL (IIIB) have short survival if not treated appropriately at early stages.

2.2 Grading criteria for neuroblastoma by WHO

Neuroblastoma is the most common extracranial solid cancer, commonly affects to infant and children (0-5 years). Based on the American Cancer Society [2] statistics, it is by far the most common cancer in infants and the third most common type of cancer in children. WHO recommends the use of the International Neuroblastoma Pathology Classification (the Shimada system) for categorization of the patients into different prognostic stages. This classification system is based on morphological characteristics of the tissue.

Figure 3 shows a relevant summary of this classification system as a tree diagram. Shimada system includes the age of patients, the degree of neuroblast differentiation, presence or absence of schwannian stormal development, mitosis and karyorrhexis index (MKI) and nodular pattern to conclude the final tissue classification as favorable histology (FH) and unfavorable histology (UH) [4, 23]. MKI index is calculated based on the number of Karyorrhectic cells per number of cells scanned in the sample (200 Karyorrhectic cells for every 5000 cells scanned).

Although the shimada system performs fine analysis in 80% cases and provides a precise decision, it may be disingenuous for heterogeneous tumors. A study by Teot et al. in 2007 [24] shows that for NB diagnosis, this

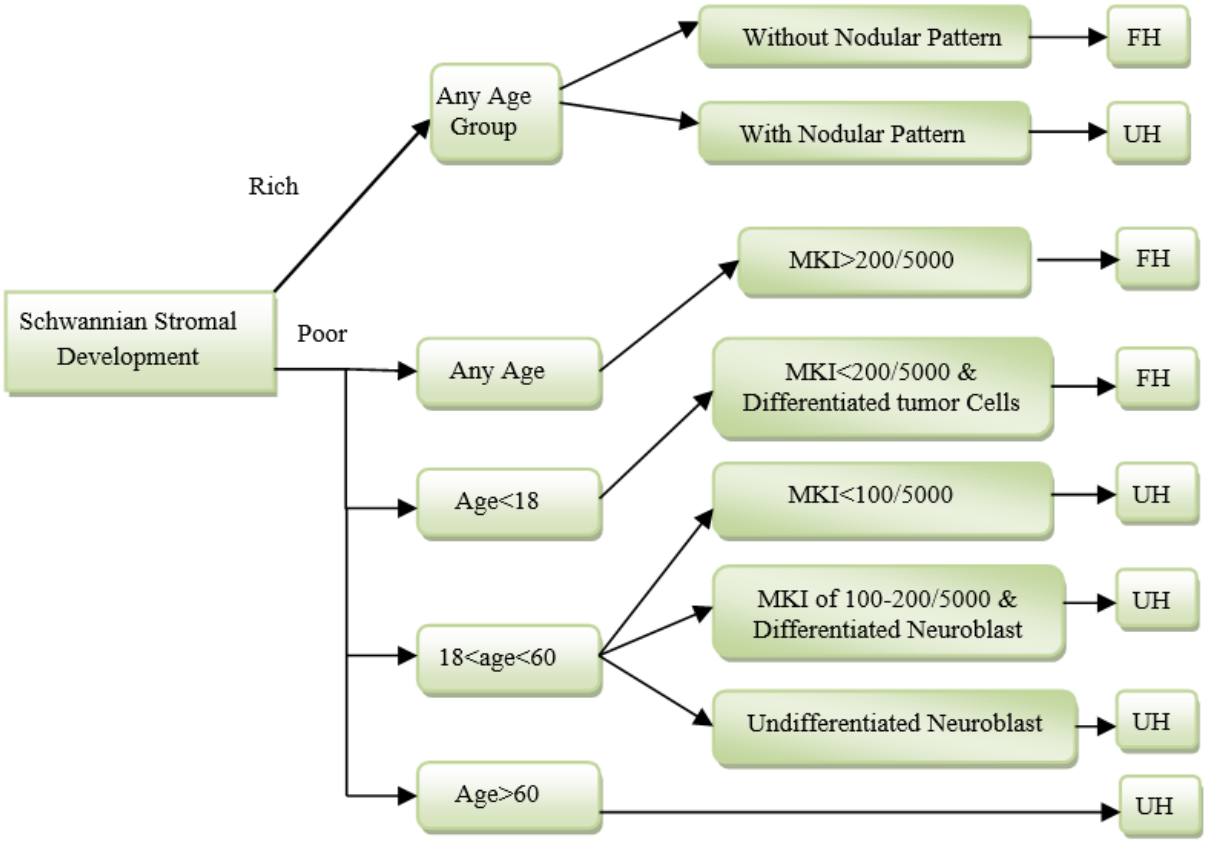


Figure 3: Shimada classification of neuroblastoma tissue sample into favorable &unfavorable histology based on schwannian stromal development.

variation can be up to 20% between central and institutional reviewers.

2.3 Gleason scale for prostate cancer

Quantitative image analysis in the context of prostate cancer is done with the help of gleason score [25], persons having high gleason score are more aggressive and have a worse prognosis. Grading is solely based on the morphological pattern of tissue samples. Dr. Gleason shows 1 to 5 histological patterns of decreasing differentiation (pattern-1 most differentiated & pattern-5 least differentiated). Based on this score, doctors predict the stage of the disease.

C. R. King et al. in 2000 [26] have found grading error (both under & over Grading) in prostate histology. Additionally, the accuracy of the classification is important to prevent making any under or over treatment. Unfortunately, this classification becomes more vulnerable when oncologist's prognostic these highly microscopic (e.g. 40x) biopsies by only picking a small representative section (e.g. 2x or 4x) and concludes for a whole slide (40x) tissue sample.

2.4 Nottingham histologic score system for breast histology

For classification of breast histology, Bloom & Richardson introduced a system in 1957 based on cellular differentiation [27]. Later, Elston-Ellis in 1991 introduced Nottingham Histologic score system grounded on three criterions: 1) How well the cancerous cell tries to recreate normal glands (amount of gland formation); 2)How ugly the cancerous cell looks (nucleolus features); 3) How much the cancerous cells are divided (mitotic activity). Pathologist gives the score to each of these criterions in 1-3 scale, and each score is added to give final total score 3-9, this final score is responsible to classify breast biopsy into 3 criterions; [28]

- Grade 1 tumors have a score of 3-5 (well differentiated, slow growing).
- Grade 2 tumors have a score of 6-7 (moderate differentiated).
- Grade 3 tumors have a score of 8-9 (poorly differentiated, highly proliferative).

New genetic re-classification of histological grades is introduced by Anna V. Ivshina et al. in 2005 due inter-observer variability among pathologists while classifying biopsy into either grade 2 or grade 3. [29]

3 Study to computer-assisted diagnosis system

3.1 Staining process of biopsy

In the current scenario, histopathological tissue analysis is done manually by visual analysis of tissue sample, obtained from the tissue biopsy of the patients. In clinical medicine, a biopsy sample is obtained from a suspicious histological section, placed onto glass slides to be examined under the microscope. Before microscopy

examination biopsy follows a series of procedure. First, step named as a fixation step, that aims to preserve the morphological and architectural structure of the sample followed by a procedure named as embedding that allows slicing the biopsy in very thin likely 2-15 μm area section. Finally, this slice goes to the staining process.

Over the several staining presents in the current scenario, Hematoxylin and Eosin (H&E stain) is one of the principles one. It is most widely used staining medical science. A combination of hematoxylin and eosin dyes produces blues, violets, and red colors to demonstrate nucleolus & cytoplasmic. It provides information about tissues and cells with a high level of detail [1]. The stain has been unchanged for many years because it works well with the variety of fixatives and displays a broad range of cytoplasmic, nuclear and extracellular matrix feature.

Another type of staining known as Immuno-Histo-Chemical (IHC) staining, is widely used in the diagnosis of abnormal cells and is used to diagnose and track specific cellular anomalies, such as cancers, by identifying those proteins that are specifically found in affected cells. In clinical practice, IHC stain is used to mark the follicle regions where B-cell are positively marked with hues of brown color as opposed to T-cell that is negatively marked with hues of gray color. Therefore, follicles, which have a higher concentration of B-cell, can be distinguished from inter-follicular regions at low magnifications (i.e., 2×,4×,8×).

3.2 Image preprocessing: color normalization

After the staining process, color normalization is very essential preprocessing step for microscopy image. This process reduces the differences in tissue samples due to variation staining conditions. Moreover, this step is to remove irrelevant tissue structures, remove noises, and enhance the contrast. Primarily preprocessing techniques aim to improve image quality for segmentation.

Color normalization step is very commonly used to lymphoma images processing. The technique helps to provide different perceptual difference among color models [8, 9, 11]. O. sertel et al. [8, 9] used the conversion from the RGB model to L*a*b*, Euclidean distance is the measure to exploit the differences while [11] used the conversion from the RGB model to the HSV model. These color models help to segmentation process by representing each cytological component with different colors. In [3] RGB color space is projected to a 1-D uni-tone image by considering only 1st PCA, to attain a single channel image that has the highest contrast.

3.3 Automated segmentation and classification of histopathological image

Use of CAD in medicine is drastically increasing day by day, to aid in detection, diagnosis of disease in digital pathology. To fulfill the requirement of the system, researchers identify certain quality parameters of images such as the morphological structure of lymphocytes, cancerous nuclei, and glands. The presence of extent shape, size and other morphological appearance of these structures are important indicators for the presence of severity of the disease. For instance, for neuroblastoma histology, the existence of a nodular pattern and distinguish and un-distinguish neuroblast leads to favorable and unfavorable histology. In terms of prostate lymphoma, the size of glands tends to reduce with higher Gleason patterns. Similarly, the presence of many lymphocytes in breast cancer histopathology is strongly suggestive of poor disease outcome and survival.

Another motivation for detecting and segmenting histological structures has to do with the need for counting of objects, generally cells or cell nuclei. Cell counts can have diagnostic significance for some cancerous conditions. For instance, in terms of FL, pathologist counts the number of CBs per HPF also in terms of NB MKI index is calculated based on the number of Karyorrhectic cells per number of cells scanned in the sample. Bibbo et al. [30] reported 1.1%-4.7% error in cell counts compared to manual counts for Feulgen-stained prostate specimens. In order to evaluate segmentation methods, several matrices have been proposed for evaluating computational and manual segmentation conducted by the oncologist. Among these matrices, common names sensitivity, specificity, accuracy considering the concept of true positive, false positive [15]. Accuracy can quantify how many pixels from manual segmentation were also identified by the computational methods [15]. Another metric named Zijdenbos similarity index [11], This measures the overlapping ration between the shapes from manual and automatic segmentation. The systematic survey is presented for diagnosing and prognosis techniques using computational techniques.

A very basic CAD model started in 2007, to aid pathologists in prognosis for histological grading system of FL histology by J. Kong et al. [5] in which pathologist's (human) intervention was also required along with computer-assisted diagnosis model in the classification of FL. Firstly, color and texture are features extracted to detect follicles from H&E stained images using K-means clustering method, followed by a morphologic post-processing step to remove the noisy regions and smooth out the boundaries of follicle regions. Subsequently, a manual registration step is performed to detect distorted of follicles and healthy tissue sample and finally classification process proceeds to group data into CB and non-CB classes. In continuation of previous work [5], O. Sertel et al. in 2008 [6] develops a computerized system to reduce manual intervention based on non-supervised clustering to assist pathologists in differentiating CB from non-CB cells. Comparative analysis between similarity index between computerized system & pathologist is discussed in table 1. Till now researchers have been only attempting to differentiate between CB and non-CB cells [5, 6], but in 2008, O. Sertel et al. [7] introduced new classification index based on WHO criteria. The WHO recommends histopathological grading of FL based on the number of CB, per standard 40x magnifications HPF of 0.159mm^2[4]. To capture this information, firstly images

were partition into distinct cytological components (nuclei, cytoplasm, RBC, background and extracellular material) based on unsupervised segmentation technique followed by color texture is extracted by non-linear quantization using self-organizing map, which is used to differentiate the low and intermediate grades (I and II) from high grades (III) of FL. For the classification of the segmented image (from the first step), principle component analysis (PCA) and linear discriminant analysis (LDA) followed by Bayesian parameter estimation are used and again O. Sertel et al. in 2009 [8], introduced a novel computer-assisted diagnosis system for digitalized histology based on model-based intermediate representation (MBIR) and incorporates texture analysis. The system identifies basic cytological component using La*b* color space if the same amount of change in color values produces the same amount of perceptual difference of visual importance. Property of La*b* color space allows to use euclidean distance in comparing colors. Clustering in La*b* color space is done using the K-means algorithm. Researcher uses PCA and FLDA to reduce the dimension, followed by bayesian classifier based on maximum a posterior decision rule for classification.

Limitations of conventional feature space-based clustering algorithm like K-mean and EM (expectation Minimization) are identified by O. Sertel et al. in 2010 [9], those are widely used for Histopathology image segmentation [5, 6, 7, 8]. Those algorithms based on prior knowledge of clusters and confined to the mostly elliptical shape of these clusters. In fact, especially in histopathology imagery, these assumptions may not suit well due to inhomogeneous intensities and overlapping cell nuclei. In order to achieve a robust approach, which can overcome the limitation of clustering-based feature space; a non-parametric method is introduced [9], namely the mean-shift method. This method is applied to L*a*b* color space and finds out stationary points of density. Adaptive thresholding technique is used to differentiate between CB and non-CB cells. In fact, this is a very initial detection system based on the non-parametric approach to distinguish CB and non-CB cells with relatively very high false positive [9].

To overcome these very high false positives is being experienced by [9], O. Sertel et al. in 2010 [3] introduces a better approach grounded on adaptive likelihood-based cell segmentation. This approach includes, Firstly, input images in the RGB color space is projected onto a 1-D uni-tone image by considering only first principal component, to attain a single channel image that has the highest contrast. This uni-tone image is further normalized between [0,1]. Subsequently, two steps process is applied for the classification of CB and non-CB cells as follows.

- Identify evident non-CB cells based on size, shape, and eccentricity.
- Redefine CB detection by learning and utilizing the texture distribution of non-CB cells.

In 2010, Siddhartha Samsi et al. [10] described an automated system to identify follicles in IHC stained tissue section rather than H&E stained images. IHC stains are typically used to identify specific categories of cells in the tissue. The proposed algorithm uses color and texture measures for identifying follicles and their respective boundarization. To reduce under and over-segmentation, the author applied the watershed segmentation algorithm and finally, fourier descriptors are used to smooth follicle boundary.

A modified paper of [3] is introduced to improve the accuracy of pathologists by extracting the morphological characteristic of objects based on novel texture features, color-space decomposition. Instead of grayscale representation of images, color-space is utilized, and RGB, L*a*b*, HSV color spaces are investigated. K. Belkacem-Boussaid et al. in 2010 [11] uses a multivariate image analysis technique using PCA to classify between CB and non-CB tissue samples. For extraction of the geometry of CB, researchers use operations such as thresholding, morphological filtering, and area identification.

Again, the different approach introduced by K. Belkacem-Boussaid et al. in 2011 [12] in which, follicular regions are identified first at lower magnification (2X). Later, higher magnification is used to identify and count the number of centroblastic cells. Three step procedures are applied to start form region-based segmentation, which is used to determine the location of the follicle regions, followed by iterative shape index calculation and finally recursive watershed algorithm is applied. For initial segmentation of follicles, Chan & Vese [13] introduced a region-based segmentation approach based on curve expansion. With respect to the interior of follicles, the curve starts moving, and energy of curve minimized where the desired boundary is achieved. During boundarization process, level set formulation leads to very closed objects (likely overlapped) to each other; hence level set formulation is venerable to overlapping combined follicles. To overcome these overlapping follicles problems, an adaptive control splitting and merging technique is applied at the object level. The study presents a control factor based on the concavity index for the individual object. Followed by a novel recursive marked-watershed operation. Combining all the operation produces a better segmentation for overlapped follicles and hence prevents limitations that usually are introduced during the traditional morphological watershed is overcome. A combined effort to reduce under and over-segmentation of overlapped cells is made by H. Kong et al. in 2011 [14]. This framework firstly segments the histopathological images based on the local context features extracted around the pixel. In which, each pixel is categorized into either the cell or extracellular classes. After that local Fourier transformation is applied for extracting the texture feature based on newly defined color space, called the most discriminant color space (MDC).

A different approach to isolating FL tissue samples, M. Oger et al. in 2012 [15] presented a noble technique. In this technique segmentation of follicular areas is recognized firstly before histological grading is done. Following the current practice of the pathologist, low resolution, lower magnification (2X) IHC Stained image uses to generate a mask of the follicular boundaries (follicular area) by a newly deployed feature-based cluster approach. Then these boundaries are map onto a

corresponding registered H&E stained image. From these H&E stained images, color and texture information is extracted by converting the image from RGB space to HSV space. Where S (saturation) channel is represented as color information. Texture information is quantified by the homogeneity of 9×9 neighborhoods of each pixel using the co-occurrence matrix approach. Resulting feature vector is classified using k-means classifiers with K=4 one for follicles, second for the intra-follicular area, third for the mixture of follicles and intra-follicular area and last for the background. With respect to previous researches [10, 15], an efficient computational framework has been proposed [16] for the analysis of whole slide images (an application to FL immunohistochemistry stain) by S. Samsi et al. in 2012. This framework involves calculation of color and grayscale features, which are used as feature vectors for K-means clustering. The color feature used is the hue channel from HSV color-space conversion of the original image. The first texture feature used is the output of a median filter of size 45 □ 45 applied to a gray-scale version of the image. The second texture feature is the energy feature calculated from the co-occurrence matrix. The output of the clustering algorithm provides a segmented image that is further processed by an iterative watershed algorithm followed by a boundary-smoothing step.

B. Oztan et al. in 2012 [17] introduced a new computer-aided technique for grading of FL tissue sample based on cell graph and multi-scale feature analysis. FL image is analyzed using the cell graph to know the structural organization of component like nuclei, cytoplasm. Cell graph technique includes a graph theory concept to represent FL image with the un-weighted and undirected graph. Nodes of the graph are represented as cell nuclei whereas adjacencies of cells are represented with the edges of the graph. After constructing the cell-graph representation, a feature space vector is formulated, which includes structural organization within the nuclei and cytoplasm components.

E. Michail et al. in 2014 [18] projected a new scheme for detection of CB from H&E Stained images. Detection of CB starts from converting the image into grayscale and filter using the Gaussian filter with kernel 3□3. Moreover, in order to detect the nuclei, the difference between nuclear membrane and background are enriched by histogram equalization technique. Later, segmentation of nuclei (dark) from extracellular material and background, otsu-thresholding is applied. Additionally, Expectation maximization (EM) technique is used to separate touching cells. Isolation is CBs is done based on shape, size, and intensity of histogram criteria. Finally, LDA is used to classify CBs and non-CBs cells.

E. N. kornaropoulosin et al. in 2014 [19] introduced a quantitative methodology to categorize FL histology into one of two categories of cells (CB vs non-CB) using linear and non-linear dimensionality reduction. In the first method, biased features are calculated with the help of singular value decomposition, in which discrimination between CB and non-CB problem can be formulated as a minimization problem, where the objective is to minimize the least-square error between given images of CB and non-CB its low-rank approximation. In the second method, the classifier is based on preserving the similarity among the images of CB and non-CB cells using Laplacian eigen-maps.

K. Dimitropoulos et al. [20] in 2014 presented a study for automatic detection of CBs from microscopic images obtained from FL biopsy. Initially, the touching-cell splitting algorithm is applied using the Gaussian mixture model and EM algorithm to segment the biopsy into basic cytological elements. Additionally, morphological and textural analysis of CBs is applied to extract various features related to the nucleolus, cytoplasm extra-cellular cells. In the final step, an innovative classification scheme is proposed based on adaptive neuro-fuzzy inference system to classify the interested cells.

Follicular lymphoma grading system, a color-coded map-based system was presented by M. Faizal Ahmad Fauzi et al. in 2015 [21].The system includes the HSV color model to register the two images to obtain a good gray level separation among the follicle regions, non-follicle regions, and the white background. Classification in respective grades has been done using KNN classifier of potential CB regions within sub-blocks of the HPF regions followed by rule-based classification at the block, HPF and tissue levels.

A novel framework is proposed by K. Dimitropoulos et al. in 2016 [22] for segmentation, separation, and classification of CBs and Non-CBs cells from H&E and IHC Stained FL histology. For segmentation of nuclei, energy minimization technique based on the graph-cut method is applied to IHC Stained images. A noble algorithm based on inspired by clustering of large-scale visual terms is used to segment the nuclei. Additionally, H&E stained images enable to extract textural information related to histological characteristics. Finally, morphological characteristic from IHC staining and textural information from H&E Stained Images are used to construct the feature vector. This feature vector is then fed into Bayesian network classifier to classify FL histology into respective grades.

In 2012, Scott Doyle et al. [31] (2012) headed a classification technique for Prostate histology. In the first step, this algorithm decomposes the whole-slide image into an image pyramid comprising multiple resolution levels. Regions identified as cancer via a Bayesian classifier at lower resolution levels are Subsequently examined in greater detail at higher resolution levels, thereby allowing for rapid and efficient analysis of large images. Later, Sahirzeeshan Ali et al. [32] in 2012, commented and enhanced classification accuracy by demonstrating a robust algorithm, which provides boundaries close to the actual nuclear boundary and the shape constraint prevents spurious edges. The algorithm also demonstrates an application of these synergistic active contour models using multiple level sets to segment nuclear and glandular structures on digitized histopathology images of breast and prostate biopsy specimens.

A Manifold learning (ML) scheme is presented by Rachel Sparks et al. [33] in 2013, which attempts to generate a low dimensional manifold representation of a

higher dimensional feature space while simultaneously preserving nonlinear relationships between object instances. Classification can then be performed in the low dimensional space with high accuracy while, Safa'a N. Al-Haj Saleh et al. [34] in 2013 uses K-means clustering based approach was employed to the a* color channel of the L*a*b* color model for each of the tissue images, followed by statistical and morphological features extraction for the segmented lumen objects and glands. Finally, a naive bayes classifier was used to classify tissue images to the correct grade. In 2016, M Khalid Khan Niazi et al. [35] provides additional view angle meaningful representation and discriminates between low and high graded prostate lymphoma. These meaningful representation features include luminal and architecture feature. These features help to create two subspaces one for prostate histology assessed as a low grade and other to classify histology into the high grade.

An idea to work with gray tone images, which are represented by two descriptors, one is local binary pattern & other is local phase quantization, a noble approach is presented by Ville Ojansivu et al. [36] in 2013. The classification of the images into the three classes was done using three one-versus-rest SVM classifiers with a radial basis function kernel (RBF) combined with a chi-square distance metric. Selecting the largest of the scores produced by the individual SVM classifiers chose the final class. Additionally, Maqlin Paramanandam et al. [37] in 2016 propose a novel segmentation algorithm for detecting individual nuclei from Hematoxylin and Eosin (H&E) stained breast histopathology images. This detection framework estimates a nuclei saliency map using tensor voting followed by boundary extraction of the nuclei on the saliency map using a loopy back propagation algorithm on a markov random field. The algorithm uses hough transformation to identify seeds points, which are used to initializing shape- and texture based active contour algorithm.

In 2009 O. Sertel et al. [23] Paper, introduced a segmentation technique grounded on texture feature, which is extracted using local binary pattern and co-occurrence. This statistical framework uses modified k-nearest neighbor classifier is used to determine the confidence level of the classification to make the decision at a resolution level.

Segmentation & Classification Techniques	Reference	Publication year	Images types	Object of Interest	Evaluation result
K-means clustering followed by manual registration	[5]	2007	FL	CB & non-CB cells classification	There is no performance evaluation for classification
Non-supervised clustering	[6]	2008	FL	CB & non-CB cells classification	Maximum accuracy of 90.7
Non-linear quantization using SOM followed by LDA & BPE	[7]	2008	FL	CBs per HPF calculation	88.9% accuracy of FL tissue classification
MBIR	[8]	2009	FL	CBs per HPF calculation	88.7% correctness achieved
Texture & color-based Classification using KNN	[23]	2009	NB	Identification of Stroma rich and stroma poor	overall classification accuracy of 88.4%.
Mean shift method	[9]	2010	FL	CB & non-CB cells classification	qualitative very high false positive appealed
Adaptive Likelihood-based cell segmentation	[3]	2010	FL	CB & non-CB cells classification	maximum accuracy 80.7% achieved
Watershed with Fourier descriptors	[10]	2010	FL	Follicular Region	87.11% accuracy while classifying follicles
Multivariate image analysis using PCA	[11]	2010	FL	CB & non-CB cells classification	82.57% precision while segmenting
Fourier transformation, most discriminant color space	[14]	2011	FL	Cellular nuclei	The total error rate is 5.25% per image
Region based segmentation followed by the iterative index	[12]	2011	FL	Follicular Region	There is no performance evaluation for classification

and finally recursive watershed					
Region-based active contours with the statistical shape model	[32]	2012	CaP & BC	Segment overlapping lymphocytes and lumen	90% accuracy while segmentation
Boosted Multi-resolution Classifier	[31]	2012	CaP	Regions of CaP	There is no performance evaluation for classification
K-Mean Classifier	[15]	2012	FL	The follicular area along with Follicles detection	There is no performance evaluation for classification
K-mean, co-occurrence matrix followed by watershed & boundary smoothing	[16]	2012	FL	Identification of follicles	There is no performance evaluation for classification
Cell graph followed by supervised learning	[17]	2012	FL	Detection of nuclei and other cytological components	87% accuracy in classification
A Statistical Shape model of manifolds	[33]	2013	CaP	Manifold regulization	There is no performance evaluation for classification
Texture Based Classification	[36]	2013	BC	Breast glands	This work achieves the classification accuracy of breast histology is 90%.
L*a*b*-based segmentation	[34]	2013	CaP	Lumen objects and tissue glands	Automated classification results achieved the accuracy of 91.66%.
Thresholding for RBC removal and Otsu to segment nuclei & LDA	[18]	2014	FL	Centro blast detection	82.58% CB were successfully detected
Linear and non-linear dimensionality reduction followed by Laplacian Eigen map classifier	[19]	2014	FL	CB & non-CB cells classification	97.67 % of accuracy
Linear and non-linear dimensionality reduction followed by orthogonal bases	[19]	2014	FL	CB & non-CB cells classification	98.22 % of accuracy
Touching cell splitting using GMM followed by neuro fuzzy inference system	[20]	2014	FL	Cytoplasmic element classification	90.35% detection rate achieved
Color coded map-based segmentation followed by rule-based classification	[21]	2015	FL	CBs per HPF calculation	80 % correct classification
Texture based Active Contour Model	[8]	2016	BC	Nuclei detection	There is no performance evaluation for classification
Luminal and Architecture based classification	[35]	2016	CaP	Nuclei detection	97.6 % accuracy while segmentation
Energy minimization using graph cut	[22]	2016	FL	CB & non-CB cells classification	94.56 % accuracy achieve

followed by Bayesian network classifier					
Entropy based histogram thresholding	[41]	2017	FL	Segmentation of CD3+ & CD3- T cells	Sensitivity and specificity measure 90.97% & 88.38% respectively

Table 1: describes the state of art CAD technologies used in Histopathology. Here results of various segmentation techniques and classification accuracy upon a different set of images pertaining to Cap, BC, FL, and NB are presented.

4 Discussion and future directions

Based on the review of the literature, it is concluded that the quantitative analysis from CAD System is very useful in decision making policies. It is successfully applied and benefitted in the diagnosis and in the classification of tissue (as FL, CaP, NB, BC) associated with various grades. The classification is often done by extracting quality parameters like nuclei, Karyorrhectic, background, glands, Centro blast, follicular region, prostate region and the extra-cellular area from the histological section of H&E Stained images. Various methodologies have been proposed to extract these quality parameters and keep visual difference based on textural heterogeneity [6, 7, 8, 10, 11, 14, 16, 19, 34, 36]. Several morphological features [8, 12, 17, 20, 22, 32, 33] and their combination with textural feature [5, 10], as well as cell graph-based features [17, 22], Otsu-thresholding [18] have been introduced to form a feature space vector. This feature vector is often identified in lower dimension using first principle component analysis [3, 7, 8, 11], which is responsible for calculating the main mode of variation in data. Reduced feature vector is then fed into various linear and non-linear classifiers (supervised clustering [17], non-supervised clustering [6], LDA followed by BPE [7, 8, 18, 22],Laplacian Eigen map classifiers [19], K-nearest neighbor [21], neuro-fuzzy inference system [20], boosted multi-resolution [31], rule-based classifiers [19]) to classify these histology into respective grades. Using these approaches classification accuracy ranges from 80% to 95% for FL sample, 90% to 97.6% for prostate samples, 90±5 % for breast histology and finally 88.4% for the neuroblastoma tissue sample. But this level of performance accuracy may not be enough in clinical application and leads to future directions;

- Due to different stain manufacturers and inconsistent biopsy staining and nonstandard imaging can cause color variation in histopathological images [40,42].
- The limited number of datasets leads to inappropriate efficiency of the proposed system in addition to that two different types of stained images (H&E or IHC) limits the robust performance of segmentation.
- Another motivation is to standardize the evaluation metrics, including accuracy. As M. Oger [15] uses specificity and sensitivity rather Zijdenbos similarity index in [11].
- The higher rate of false positive is still a challenge for FL cases [9].

Over a few decades, a lot of segmentation and classification techniques have been introduced, every researcher has the goal to classify the histopathological images into their respective grades. The essential quality parameter, which can represent these histopathological images, is usually a pre-requisite regarding the classification and state of disease. Other important parameters can be a choice of the classifier, which can deal with large and highly dense datasets. So, this study shows few classifiers and their significance regard to disease grading. Classification accuracy with difference classifier and the quality metrics extracted from histopathological images is shown in Table 1.

In the case of FL, histopathological grading is based on the number of large malignant cells named as CB. It is inferred from the analysis that the demarcation of CB can improvise the classification accuracy of grades. Moreover, merging of two regions should be minimized in order to compute the disease grading with the desired accuracy. Similarly, NB counting the karyorrhectic cell along with differentiated architecture of nodular pattern leads segmentation accuracy 88.4%. On the other hand, prostate histology, which is solely based on morphological pattern, better computer-assisted view angle color descriptor based meaningful representation leads to max 97.6% accuracy for segmentation. In case of breast histology, proper segmentation of glands and their representation into biopsy using local binary pattern and local phase quantization, and subsequent classification with SVM classifier leads to more precise segmentation accuracy of 90%.

5 Reference

[1] M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot and B. Yener, “Histopathology Image Analysis: A Review” IEEE Rev. Biomedical Engineering, vol. 2, pp.147-171, (2009). https://doi.org/10.1109/RBME.2009.2034865

[2] L. R. Teras, Carol E. DeSantis, James R. Cerhan, Lindsay M. Mortan, A. Jemal, Christopher R. Flowers, “2016 US Lymphoid Malignancy Statistics by World Health Organization Subtypes”, CA: Cancer, Vol. 66, no. 6, (2016). https://doi.org/10.3322/caac.21357

[3] O. Sertel, G. Lozanski, A. Shana’ah, and M. N. Gurcan, “Computer-aided Detection of Centro blast for Follicular Lymphoma Grading using Adaptive Likelihood based Cell Segmentation ” IEEE Trans Biomed Engineering pp. 2613–2616, (2010). https://doi.org/10.1109/TBME.2010.2055058

[4] S. Swerdlow, E. Campo, N. Harris, E. Jaffe, S. Pileri, Stein, H. Thiele, and J. Vardiman, “WHO classification of tumors of hematopoietic and

lymphoid tissues," vol. 2, World Health Organization, Lyon, France, fourth ed. 117(19): 5019–5032 (2008).10.1182/blood-2011-01-293050.

[5] J. Kong, O. Sertel, A. Gewirtz, A. Shana'ah, F. Racke, J. Zhao, K. Boyer, U. Catalyurek, M. N. Gurcan, G. Lozanski, "Development of computer based system to aid pathologists in histological grading of follicular lymphomas", GA. American Society of Histology (2007). https://doi.org/10.1182/blood.V110.11.3318.3318

[6] O. Sertel, J. Kong, G. Lozanski, U. Catalyurek, J. H. Saltz, Metin N. Gurcan, "Computerized microscopic image analysis of follicular lymphoma", SPIE vol. 6915, Medical Imaging (2008). https://doi.org/10.1117/12.770936

[7] O. Sertel, J. Kong, U. Catalyurek, G. Lozanski, A Shanaah, J. H. Saltz, M. N. Gurcan, "Texture classification using nonlinear color quantization: Application to histopathological image analysis" IEEE ICASSP'08; Las Vegas, NV (2008). https://doi.org/10.1109/ICASSP.2008.4517680

[8] O. Sertel, J. Kong, U. Catalyurek, G. Lozanski, Joel H. Saltz, Metin N. Gurcan, "Histopathological Image Analysis Using Model-Based Intermediate Representations and Color Texture: Follicular Lymphoma Grading", Journal Signal Process System, vol. 55, pp.169–183 (2009). https://doi.org/10.1007/s11265-008-0201-y

[9] O. Sertel, U. Catalyurek, G. Lozanski, A. Shanaah, Metin N. Gurcan, "An Image Analysis Approach for Detecting Malignant Cells in Digitized H&E-stained Histology Images of Follicular Lymphoma" International Conference on Pattern Recognition (2010). https://doi.org/10.1109/ICPR.2010.76

[10] S. Samsi, G. Lozanski, A. Shana'ah, M. N. Gurcan, "Detection of Follicles from IHC Stained Slide of Follicular lymphoma Using Iterative Watershed", IEEE transaction Biomedical Eng. pp. 2609-2612 Oct. (2010). https://doi.org/10.1109/TBME.2010.2058111

[11] K. Belkacem-Boussaid, M. Pennell, G. Lozanski, A. Shana'ah, and M. Gurcan, "Computer-aided classification of centroblast cells in follicular lymphoma", Anal. Quant. Cytol. Histol., vol.32 no. 5, pp. 254–260 (2010). PMC ID: [PMC3078581].
K. Belkacem-Boussaid, S. Samsi, G. Lozanski, M.N. Gurcan, "Automatic detection of follicular regions in H&E images using iterative shape index", Computerized Medical Imaging and Graphics 35 pp. 592–602, (2011). https://doi.org/10.1016/ j.compmedimag.2011.03.001

[12] T. F. Chan, Vese LA. Active contours without edges. IEEE Transaction Image Processing, vol. 10, (2001). https://doi.org/10.1109/83.902291

[13] H. Kong, M.N. Gurcan, and K. Belkacem-Boussaid, "Partitioning Histopathological Images: An Integrated Framework for Supervised Color-Texture Segmentation and Cell Splitting" IEEE Transactions On Medical Imaging, Vol. 30, No. 9, (2011). https://doi.org/10.1109/TMI.2011.2141674

[14] M. Oger, Philippe Belhomme, Metin N. Gurcan, "A general framework for the segmentation of follicular lymphoma virtual slides" Computerized Medical Imaging and Graphics vol. 36, pp. 442–451 (2012). https://doi.org/10.1016/j.compmedimag.2012.05.00 3

[15] S. Samsi, Ashok K. Krishnamurthy, Metin N. Gurcan, "An efficient computational framework for the analysis of whole slide images: Application to follicular lymphoma immunohistochemistry", Journal of Computational Science vol. 3, pp. 269–279 (2012). https://doi.org/10.1016/j.jocs.2012.01.009

[16] B. Oztan, H. Kong, M. N. Gurcan, & B. Yener, "Follicular Lymphoma Grading using Cell-Graphs and Multi-Scale Feature Analysis", Medical Imaging, Proc. of SPIE Vol. 8315 (2012). https://doi.org/10.1117/12.911360

[17] E. Michail, Evgenios N. Kornaropoulos, Kosmas Dimitropoulos, Nikos Grammalidis, Triantafyllia Koletsa, Ioannis Kostopoulos, "Detection of Centro blasts in H&E Stained Images of Follicular Lymphoma" 2014 IEEE 22nd Signal Processing and Communications Applications Conference 2319-2322 (2014). https://doi.org/10.1109/SIU.2014.6830728

[18] E. N. Kornaropoulos, M Khalid Khan Niazi, Gerard Lozanski, and Metin N. Gurcan, "Histopathological image analysis for centroblasts classification through dimensionality reduction approaches", Cytometry Analysis, vol. 85, no.3: 242–255 (2014). https://doi.org/10.1002/cyto.a.22432

[19] K. Dimitropoulos, E. Michail, T. Koletsa, I. Kostopoulos, N. Grammalidis, "Using adaptive neuro-fuzzy inference systems for the detection of centroblasts in microscopic images of follicular lymphoma", Signal, Image Video Process, 8 (1), pp. 33–40, (2014). https://doi.org/10.1007/s11760-014-0688-6

[20] M. F. A. Fauzi, M. Pennell, B. Sahiner, W. Chen, A. Shana'ah, J. Hemminger, A. Gru, H Kurt, M. Losos, A. Joehlin-Price, C. Kavran, S. M. Smith, N. Nowacki, S. Mansor, G. Lozanski and Metin N. Gurcan, "Classification of follicular lymphoma: the effect of computer aid on pathologists grading", BMC Medical Informatics and Decision Making, vol. 15 (2015). https://doi.org/10.1186/s12911-015-0235-6

[21] K. Dimitropoulos, P. Barmpoutis, T. Koletsa, I. Kostopoulos, N. Grammalidis, "Automated detection and classification of nuclei in pax5 and H&E-stained tissue sections of follicular lymphoma" Signal, Image Video Process, vol.11, no. 1, pp. 145–153(2016). https://doi.org/10.1007/s11760-016-0913-6

[22] O. Sertel, J. Kong, H. Shimada, U.V. Catalyurek, J.H. Saltz, &M.N. Gurcan, "Computer aided Prognosis of Neuroblastoma on Whole-slide Images: Classification of Stromal Development", Pattern Recognit. vol. 42, no. 6, pp. 1093– 1103, (2009).

https://doi.org/10.1016/j.patcog.2008.08.027

[23] LA. Teot, RSA. Khayat, S. Qualman, G. Reaman, D. Parham, “The problem and promise of central pathology review: Development of a standardized procedure for the children’s oncology group”, Pediatric and Developmental Pathology, pp. 199–207, (2007). https://doi.org/10.2350/06-06-0121.1

[24] D. Gleason, “Classification of prostatic carcinomas”, cancer chemother Rep. pp. 125-128, (1966). [PubMed ID: 5948714].

[25] C. R. King, JP. Long, “prostate biopsy grading error: a sampling problem?”, International Journal cancer, (2000). [Pubmed ID: 11180135].

[26] H. J. Bloom, W. Richardson, "Histological grading and prognosis in breast cancer; A study of 1409 cases of which 359 have been followed for 15 years" British Journal of Cancer, Vol. 11 no. 3, pp. 359–377, (1957). https://doi.org/10.1038/bjc.1957.43

[27] CW. Elston, IO. Ellis, “Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up”. Histopathology, vol. 19: pp. 403–410, (1991). https://doi.org/10.1111/j.1365-2559.1991.tb00229.x

[28] A. V. Ivshina, J. George, O. Senko, B. Mow, T. C. Putti, J. Smeds, T. Lindahl, Y. Pawitan, P. Hall, H. Nordgren, J. E.L. Wong, E. T. Liu, J. Bergh, V. A. Kuznetsov and L. D. Miller, “Genetic Reclassification of Histologic Grade Delineates New Clinical Subtypes of Breast Cancer” Cancer research, vol. 21, (2006). https://doi.org/10.1158/0008-5472.CAN-05-4414

[29] Bibbo M, Kim DH, Pfeifer T, Dytch HE, Galera-Davidson H, Bartels PH. “Histometric features for the grading of prostatic carcinoma”, Anal Quant Cytol. Histol. vol. 13, pp. 61–68 (1991). [PMID: 2025375].

[30] S. Doyle, M. Feldman, J. Tomaszewski, & A. Madabhushi, “A Boosted Bayesian Multiresolution Classifier for Prostate Cancer Detection from Digitized Needle Biopsies”, IEEE trans. on biomedical eng. vol. 59, no. 5, (2012). https://doi.org/10.1109/TBME.2010.2053540

[31] S. Ali & A. Madabhushi “An Integrated Region-, Boundary-, Shape-Based Active Contour for Multiple Object Overlap Resolution in Histological Imagery” IEEE Transaction on Medical Imaging vol. 31, no. 7, pp. 1448-1460, (2012). https://doi.org/10.1109/TMI.2012.2190089

[32] R. Sparks, A. Madabhushi, “Statistical shape model for manifold regularization: Gleason grading of prostate histology”, Computer Vision and Image Understanding, pp.1138-1146, (2013). https://doi.org/10.1016/j.cviu.2012.11.011.

[33] Safa’a N. Al-Haj Saleh, Moh'd B. Al-Zoubi, “Histopathological Prostate Tissue Glands Segmentation for Automated Diagnosis”, 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, (2013). https://doi.org/10.1109/AEECT.2013.6716471.

[34] M. K. Khan Niazi, K. Yao, D. L Zynger, S. K Clinton, J. Chen, M. Koyutürk, T. La. Framboise, M. Gurcan, “Visually Meaningful Histopathological Features for Automatic Grading of Prostate Cancer”, IEEE Journal of Biomedical and Health Informatics, pp. 2168-2194, (2016). doi: 10.1109/JBHI.2016.2565515.

[35] V. Ojansivu, N. Linder, E. Rahtu, M. Pietikainen, M. Lundin, H. Joensuu, J. Lundin, “Automated classification of breast cancer morphology in histopathological images”, Diagnostic Pathology, (2013). https://doi.org/10.1186/1746-1596-8-S1-S29

[36] M. Paramanandam, M. O’Byrne, B. Ghosh, J. J. Mammen, M. T. Manipadam, R. Thamburaj, V. Pakrashi, “Automated Segmentation of Nuclei in Breast cancer Histopathology Images”, PLOS ONE (2016). https://doi.org/10.1371/journal.pone.0162053

[37] G. Anneke, B. Bouwer, G. W. Imhoff, R. Boonstra, E. Haralambieva, A. Berg, B. Jong, “Follicular Lymphoma grade 3B includes 3 cytogenetically defined subgroups with primary t(14;18), 3q27, or other translations: t(14;18) and 3q27 are mutually exclusive” blood journal hematology library, Feb. (2013). https://doi.org/10.1182/blood.V101.3.1149

[38] Robert Kridel, Laurie H. Sehn, Randy D. Gascoyne, “Predicting and preventing Transformation of Follicular Lymphoma”, Blood Journal ed .1. (2017). https://doi.org/10.1182/blood-2017-05-786178

[39] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, ‘Structure preserving color normalization and sparse stain separation for histological images”. IEEE Trans. Med, Imaging; vol. 35, no. 8, pp. 1962–1971 (2016). https://doi.org/10.1109/TMI.2016.2529665

[40] F. S. Abas, A. Shana'ah, B. Christian, R. Hasserjian, A. Louissaint, M. Pennell, B. Sahiner, W. Chen, M. K. K. Niazi, G. Lozanski, M. Gurcan, “Computer-assisted quantification of CD3+ T cells in follicular lymphoma”, International Society for Advancement of Cytometry, vol. 91, no. 6, pp. 609-621, (2017). https://doi.org/10.1002/cyto.a.23049

[41] A. Thaína, T. Azevedo, A. Leandro, C. Neves, Z. Marcelo, “Segmentation methods of H&E-stained histological images of lymphoma: A review,” Informatics in Medicine, 35–43, (2017). https://doi.org/10.1016/j.imu.2017.05.009

Decision Tree Algorithm Based University Graduate Employment Trend Prediction

Fen Yang
Henan University of Animal Husbandry and Economy, Zhengzhou, Henan 450044, China
E-mail: yfh508@126.com

Student paper

Keywords: data mining, decision tree, employment prediction, C4.5 algorithm

Received: November 21, 2018

The employment situation of college graduates is becoming more and more serious. It is of great significance to find effective methods to predict the employment trend of students. In this study, C4.5 algorithm was used to predict the employment trend of students. Taking the 2016 graduates of Henan University of Animal Husbandry and Economy as examples, four attributes affecting employment units were extracted, the information gain rate was calculated, the decision tree was constructed, and the classification rules were obtained. After data collection, conversion and cleaning, 420 employment records were obtained; 320 records were taken as the training samples. The classification rules were tested using 100 experimental samples, and the accuracy rate was 81%. Finally, the employment trend of the 2018 graduates was predicted by C4.5 algorithm, which provides a theoretical guidance for the arrangement of employment work in schools. Predicting the employment trend of students with decision tree algorithm is feasible and of great significance to the employment guidance of schools and the employment choice of students.

Povzetek: V tem študentskem prispevku je bil uporabljen algoritem C4.5 za analizo zaposlitve študentov na Kitajskem.

1 Introduction

The employment problem has gained more and more widespread social concern [1]. In recent years, the number of university graduates is increasing every year, and the employment situation is becoming more and more serious. It is very important for schools to analyze and study the information about students' employment, which can help them train students according to the market demand [2]. However, with the increase of the number of students, the data of employment information of graduates are accumulating continuously [3], which brings great difficulties to employment analysis. With the progress of science and technology, many new technologies have been applied in employment analysis. Wang [4] combined the residual modified GM (1,1) model with the improved neural network to predict the employment information index of graduate students, so as to predict the employment trend of graduate students. He found that the mean square error decreased from 10-1 to 10-5 with the progress of training and the performance of the algorithm was the best when the gradient value and learning rate were 7.5912×10-5 and 0.8421. Liu et al. [5] proposed a method of information gain with weight based decision tree. The weighed based information gain was obtained by genetic algorithm. The decision tree was constructed and tested on undergraduates. They found that the method had a favourable prediction accuracy. Kwak et al. [6] found that education and gender were the most important factors affecting the employment of young and middle-aged people, while gender, health status and education were the most important factors affecting the elderly. Tan et al. [7] made the short-term employment forecast of Shandong province through independent component analysis (ICA) and found that the quality of labor force, industrial structure and income were the most important factors affecting employment. Decision tree algorithm shows a good performance in data prediction. Daga et al. [8] predicted high-risk renal transplantation using decision tree and random forest and found that the accuracy rate reached 85%, which provides accurate decision support for doctors. Mohreji et al. [9] combined with the decision tree algorithm to predict the delay of air transportation. Through the study of three New York airports, it was found that the confidence level of the prediction results was very high in at least 70% of the time. At present, most of the researches on employment trend focus on the influencing factors of employment, while few researches focus on the accurate prediction of employment trend, and the traditional data processing methods are difficult to extract useful information from the historical employment data. Therefore, in this study, C4.5 algorithm from

decision tree algorithm was used to extract four decision attributes from the employment-related information of the 2016 graduates of Henan University of Animal Husbandry and Economy, and classification rules were extracted for employment trend prediction, so as to study the reliability of this method in employment prediction.

2 Employment trend forecast and data mining

Under the influence of population growth, popularization of higher education and expansion of enrollment, the number of college students has increased explosively, the number of graduates has risen sharply every year, and the employment situation has become more and more serious [10], which has aroused widespread concern of the society. The employment situation of college students is closely related to the future construction of schools, students' personal development and social stability [11]. The growth of graduates' employment rate can lead to economic growth [12]. In order to effectively manage the employment situation of students, colleges and universities have adopted information technology to collect and manage the employment information of students, in order to obtain valuable information, analyze the factors affecting employment, and help improve the employment situation. However, with the growth of the number of students, the data in the information management system is also growing rapidly. Traditional information analysis methods can not deal with such a large amount of information, nor can they fully play the potential value of these data. Although there are enough data, it is impossible to obtain the implicit association and rules between the data [13] and predict the future employment development based on these information. Therefore, an intelligent and reliable method is urgently needed to solve this problem.

Data mining technology can process massive information quickly and efficiently and extract valuable information from it. It has a wide range of applications in fields such as business, industry and military. Mining and analyzing the employment information of graduates through data mining technology to obtain the factors affecting employment can help the school employment guidance center to guide the employment of students and promote employment. It can also predict the employment trend of graduates based on the information, so as to provide a decision-making basis for the adjustment of school teaching and employment work.

3 Decision tree algorithm

Decision tree algorithm is a typical technology of data mining. It can obtain valuable information by concluding and classifying data based on the attributes of data [14]. Applying decision tree algorithm in the analysis of employment information and obtaining relevant information affecting employment through construction of decision tree and extraction of classification rules is effective in predicting the future employment trend [15]. C4.5 algorithm from decision tree was used to process and analyze employment information.

C4.5 algorithm is an improvement of ID3 algorithm [16], which selects the node attribute of tree based on information gain ratio. Data set K was defined, including k data samples, and its class attribute was set as m values, corresponding to m categories $C_i(i=1,2,\cdots,m)$. If k_i refers to the number of samples in category C_i, the amount of information needed by classification of a given object was called the entropy before division of K, and its computational formula was:

$$I(k_1,k_2,\cdots,k_m)=-\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

where p_i ($p_i=\frac{k_i}{k}$) stands for the probability of a data object belonging to category C_i. Entropy can reflect the average uncertainty and purity of data set. The larger the value of entropy, the higher the average uncertainty and the lower the purity.

Suppose that there were w discrete attribute values in attribute A and set K was divided into w subsets, $\{k_1,k_2,\cdots,k_w\}$, and the samples in K_j had the same values in attribute A, $a_j(j=1,2,\cdots,w)$. When A was taken as the testing attributes, these subsets were corresponding to some branch which grew from the node which contained set K. Suppose k_{ij} as the number of sample in subset K_j belonging to category C_i, then attribute A was divided into the entropy of subset:

$$E(A)=-\sum_{j=1}^{w}\frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s} I(s_{1j},s_{2j},\cdots,s_{mj}) \quad (2)$$

The information gain of attribute A was:

$$Gain(A)=I(s_1,s_2,\cdots,s_m)-E(A) \quad (3)$$

The larger the information gain in the set, the higher the purity of subset division.

The information gain ratio of attribute A was:

$$GainRatio(A)=\frac{Gain(A)}{SplitInfo(A)}, \quad (4)$$

where split information

$$SplitInfo(A)=-\sum_{j=1}^{w}\frac{k_{1j}+k_{2j}+\cdots+k_{mj}}{k}\log_2\frac{k_{1j}+k_{2j}+\cdots+k_{mj}}{k}$$

represents the span and uniformity of split data set K of attribute A.

4 Decision tree algorithm based employment trend prediction

4.1 Data collection and preprocessing

The 2016 graduates of Henan University of Animal Husbandry and Economy were taken as research subjects. The basic information, achievement information and employment information of the students were obtained from the student status management system, student learning management system and student employment management system, and 500 records were selected as samples.

There were many duplicate data or blank parts in the obtained data set, and the form of data was also not unified; hence preprocessing was needed.

(1) Data integration:
The data exported from the three systems were integrated into a table of general information, and the attributes are shown in Table 1.

Name	Major
Gender	Academic performance
Politics status	English competence
Student cadre	Computer skills
Participation in student society	Employment unit

Table 1: General information.

(2) Data correlation analysis:
There were many irrelevant information in the data derived from the three systems, such as name, gender, politics status, student cadres, and participation in student society, which needed to be eliminated.

(3) Data conversion:
Noise was eliminated from the data. In order to facilitate statistics and analysis, it was necessary to generalize the remaining five attributes, i.e., divide major into three categories, popular, general and unpopular, divide academic performance into excellent, general and poor, divide the English competence into CET4 and above and below CET4, divide computer skills into level 3 and above and below level 3, and divide employment unit into state-owned enterprise, private enterprise and others, represented by A, B and C.

(4) Data cleaning:
Duplicate data and blank data were deleted from the data, and finally 420 records were obtained, 320 of which were used as training samples and the remaining 100 was used for testing.

4.2 Establishing decision tree

The training samples were analyzed by taking employment unit (A, B and C) as the labeling attribute and major, academic performance, English competence and computer skills as decision attributes. The number of students under different categories of different attributes is shown in Table 2.

Decision-making attribute		A	B	C
Major	Popular	43	21	17
	General	12	38	66
	Unpopular	3	47	73
Academic performance	Excellent	21	17	50
	General	28	34	67
	Poor	18	26	59
English competence	CET4 and above	46	52	61
	Below CET4	36	46	79
Computer skills	Level 3 and above	56	49	61
	Below level 3	62	58	34

Table 2: Training data set.

Suppose training sample as K and the corresponding subset and number of A, B and C as $k_1 = 55, k_2 = 158, k_3 = 107$ respectively. The entropy of K was calculated using equation (1).

$$I(k_1, k_2, k_3) = I(55,158,107) = 2.315878$$

Then the information gains of different decision attributes were calculated using equation (2), (3) and (4).

(1) Major
Major was divided into popular, general and unpopular. When the major was popular, the entropy was:

$$I(43,32,27) = -\frac{43}{81}\log_2\frac{43}{81} - \frac{32}{81}\log_2\frac{32}{81} - \frac{27}{81}\log_2\frac{27}{81} = 1.2145$$

When the major was general, then the entropy was:

$$I(12,38,66) = -\frac{12}{116}\log_2\frac{12}{116} - \frac{38}{116}\log_2\frac{38}{116} - \frac{66}{116}\log_2\frac{66}{116} = 1.4511$$

When the major was unpopular, the entropy was:

$$I(3,47,73) = -\frac{3}{123}\log_2\frac{3}{123} - \frac{47}{123}\log_2\frac{47}{123} - \frac{73}{123}\log_2\frac{73}{123} = 0.9512$$

then the entropy of attribute "major" was:

$$E(major) = 1.5421$$

the information gain was:

$$Gain(major) = 0.0215$$

information gain rate was:

$$GainRatio(major) = 0.0131$$

(2) Academic performance
The academic performance was divided into excellent, general and poor. When the academic performance was excellent, the entropy was:

$$I(21,17,50) = -\frac{21}{88}\log_2\frac{21}{88} - \frac{17}{88}\log_2\frac{17}{88} - \frac{50}{88}\log_2\frac{50}{88} = 1.2356$$

When the academic performance was general, the entropy was:

$$I(28,34,67) = -\frac{28}{129}\log_2\frac{28}{129} - \frac{34}{129}\log_2\frac{34}{129} - \frac{67}{129}\log_2\frac{67}{129} = 1.4201$$

.

When the academic performance was poor, the entropy was:

$$I(18,26,59) = -\frac{18}{103}\log_2\frac{18}{103} - \frac{26}{103}\log_2\frac{26}{103} - \frac{59}{103}\log_2\frac{59}{103} = 1.2014$$

,

then the entropy of attribute academic performance was:

$$E(\text{subject - specific achievement}) = 1.4058$$

,

the information gain was

$$Gain(\text{subject - specific achievement}) = 0.0289$$

,

the information gain ratio was:

$$GainRatio(\text{subject - specific achievement}) = 0.0134$$

.

(3) English competence

English competence was divided into CET4 and above and below CET4. When English competence was above CET4, the entropy was:

$$I(46,52,61) = -\frac{46}{159}\log_2\frac{46}{159} - \frac{52}{159}\log_2\frac{52}{159} - \frac{61}{159}\log_2\frac{61}{159} = 0.8412$$

.

When English competence was below CET4, the entropy was:

$$I(36,46,79) = -\frac{36}{161}\log_2\frac{36}{161} - \frac{46}{161}\log_2\frac{46}{161} - \frac{79}{161}\log_2\frac{79}{161} = 1.0258$$

,

then the entropy of attribute English competence was:

$$E(\text{English competence}) = 1.3025$$

,

information gain was:

$$Gain(\text{English competence}) = 0.2157$$

,

information gain ratio was:

$$GainRatio(\text{English competence}) = 0.1656$$

.

(4) Computer skills

Computer skills was divided into level 3 and above and below level 3. When computer skills was level 3 or above, the entropy was:

$$I(56,49,61) = -\frac{56}{166}\log_2\frac{56}{166} - \frac{49}{166}\log_2\frac{49}{166} - \frac{61}{166}\log_2\frac{61}{166} = 0.9785$$

,

when computer skills was below level 3, the entropy was:

$$I(62,58,34) = -\frac{62}{154}\log_2\frac{62}{154} - \frac{58}{154}\log_2\frac{58}{154} - \frac{34}{154}\log_2\frac{34}{154} = 1.6124$$

,

then the entropy of attribute competence skills was:

$$E(\text{computer skills}) = 1.4275$$

,

the information gain was:

$$Gain(\textit{computer}\ \text{competence}) = 0.2144$$

,

the information gain rate was:

$$GainRatio(\text{computer skills}) = 0.1502$$

.

It was found from the above calculation results that the information gain rate of English competence was the largest. Therefore the attribute was regarded as the root node of decision tree. Then the information gain rate of every subtree was calculated according to the above procedures. Finally the decision tree in Figure 1 is obtained.

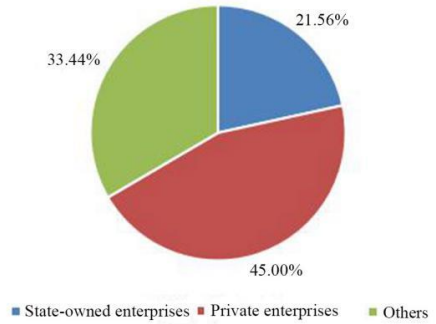


Figure 2: The prediction of employment trend of the 2018 graduates.

4.3 Generating classification rules

According to the decision tree in Figure 1, the following classification rules were obtained.

(1) If English competence = CET 4 and above AND Computer skills = level 3 and above AND academic performance = excellent AND major = general Then employment unit = state-owned enterprise

(2) If English competence = CET 4 and above AND Computer skills = below level 3 AND academic performance = excellent AND major = general Then employment unit = state-owned enterprise

(3) If English competence = CET 4 and above AND Computer skills = below level 3 AND academic performance = general AND major = general Then employment unit = private enterprise

(4) If English competence = CET 4 and above AND Computer skills = below level 3 AND academic performance = poor AND major = unpopular Then employment unit = private enterprise

(5) If English competence = below CET4 AND computer skills = level 3 and above AND academic performance = excellent AND major = popular Then employment unit = state-owned enterprise

(6) If English competence = below CET4 AND computer skills = level 3 and above AND academic performance = general AND major = general Then employment unit = private enterprise

(7) If English competence = below CET4 AND computer skills = level 3 and above AND academic performance = excellent AND major = general Then employment unit = private enterprise

(8) If English competence = below CET4 AND computer skills = below level 3 AND academic performance = poor AND major = unpopular Then employment unit = others

(9) If English competence = below CET4 AND computer skills = below level 3 AND academic performance = poor AND major = unpopular Then employment unit = others

It was concluded from the above classification rules that English competence and computer skills had the

greatest impact on the employment units of students. Students with good English competence and excellent computer skills generally worked in state-owned or private enterprises, while students with poor English competence and weak computer skills, except for some students who had good academic performance or were major in popular subjects, did not work in the state-owned enterprises or private enterprises, which showed that schools need to strengthen the training of English and computer skills and pay more attention to these two aspects in the arrangement of teaching work and students themselves should strive to improve their English and computer skills and strengthen their competitive advantage in employment.

4.4 Testing of classification performance

The effectiveness of classification rules was tested through 100 experimental samples. Then the results were compared with the actual employment unit of students. The testing results are shown in Table 3.

Sample	Classification results	Actual results
1	A	A
2	B	B
3	A	A
4	C	C
5	B	C
6	B	B
7	B	B
……		
99	B	B
100	C	C

Table 3: The testing results of classification rules.

The classification results of 81 samples were the same with the actual conditions, and the classification of 19 samples was wrong; the accuracy rate was 81%. It indicated that the obtained classification rules were relatively accurate and could determine the employment condition of students.

4.5 Prediction of employment trend

After verifying the accuracy of the classification rules, the employment trend of graduates was predicted using the method proposed in this study. The 2018 graduates were taken as examples. The information about the major, academic performance, English competence and computer skills of the students were exported from the student status management system and the student learning management system. Then the employment trend of the graduates was predicted. The results are shown in Figure 2.

Figure 2 demonstrates that the number of students who may be employed in private enterprises was the largest, accounting for 45%, while the number of students who may be employed in state-owned enterprises was the lowest, accounting for 21.56%. The decision tree and classification rules in this study could make a good prediction on the employment trend of graduates, help schools efficiently find the future employment direction of students, provide a strong basis for student employment guidance, and offer schools with valuable information.

5 Discussion

Employment has always been a problem that is difficult to be ignored and also can not be ignored in modern society, especially among university graduates. With the increase of the number of graduates, employment competition is becoming more and more fierce [17]. Employment is the most serious and difficult problem for graduates after they leave school and enter society, and it is also very important for schools. At present, all universities have employment guidance centers to collect and analyze the employment situation of students in order to find some rules and forecast employment. Employment prediction has great significance for graduates' employment and school teaching work [18]. However, with the increase of the number of university students and the accumulation of data, the analysis and processing of employment information is becoming more and more difficult, and it is difficult to obtain valuable information from mass data.

The development of data mining technology has brought about new changes. Decision tree algorithm is an efficient classification method, and it is also applicable in the prediction of employment trend. In this study, C4.5 algorithm which was relatively mature was selected. After obtaining the relevant data and information of graduates, four decision-making attributes, major, academic performance, English competence and computer skills, were extracted for analysis of employment units. The decision tree was established step by step after the calculation of information gain ratio of the attributes, and then classification rules were obtained through the decision tree.

The information gain ratio of major, academic performance, English competence and computer skills was 0.0131, 0.0134, 0.1656 and 0.1502, respectively. It was found that English competence and computer skills were the most important factors affecting the employment of graduates. In the process of employment, English competence and computer skills are the signs of graduates' ability. Many employers have specific requirements for the English and computer skills of employees. At present, schools have attached great importance to the cultivation of students' abilities in these two aspects. The extensive arrangement of English courses and computer courses has promoted the improvement of students' abilities to a certain extent. Under the rigid requirements, they have to strengthen the study of these two aspects. However, passive learning is not enough. The importance of English and computer skills must be fully recognized, which can be fully illustrated by classification rules. The extraction

of classification rules can help schools and students clearly understand what ability is the most important and crucial. On the one hand, it is conducive to the arrangement of school teaching and employment guidance; on the other hand, it is also conducive to students' active learning.

The testing of classification rules suggested that the classification rules obtained in this study had an accuracy rate of 81%, which showed that this method was feasible in predicting the employment trend of graduates. It was found from the employment trend prediction results of the 2018 graduates that many students will be employed by private enterprises and few students will be employed by state-owned enterprises. It indicated that schools should strengthen the output of talents to state-owned enterprises and carry out targeted talent training.

This paper preliminarily discussed the role of decision trees in college students' employment trend prediction, but there are still some problems that need further research:

(1) more detailed division of employment units for college students is needed;
(2) more factors that can affect college students' employment should be considered, such as family conditions, personal strengths, etc. For example, literature [19] points out that gender also can affect students' employment choices;
(3) the possibility of the application of more data mining algorithms in the employment trend prediction of college students should be analyzed. For example, the Bayesian algorithm was used for employment prediction in literature [20].

6 Conclusion

The decision tree algorithm can help handle and analyze the employment situation of students and understand the main factors affecting the employment of students. This study constructed the decision tree and extracted the classification rules through the four decision attributes, major, academic performance, English competence and computer skills. It was found that English competence and computer ability had the greatest impact on students' employment. The test suggested that the classification rules in this study had an accuracy of 81% and was feasible in predicting the employment trend of graduates. There are many shortcomings in this study. For examples, more decision attributes which can affect the employment units of students can be mined, employment units can be further divided to obtain more detailed employment trend, and a larger sample size is needed for determining the accuracy of the method.

7 Acknowledgement

This study was supported by the Research Project of Humanities and Social Sciences of Education Office of Hubei under grant number 16Z015.

8 References

[1] Xia X, Liu J (2014). Genetic Algorithm Based Forecasting Model for the Employment Demand of Major in English. International Conference on Intelligent Systems Design & Engineering Applications, pp. 331-335.

[2] Rizal MT, Yusof Y (2017). Application of data mining in forecasting graduates employment. Journal of Engineering & Applied Sciences, 12(16), pp. 4202-4207. https://doi.org/10.3923/jeasci.2017.4202.4207

[3] Miao S, Zuo J (2013). The Data Mining Technique Based on The Decision Tree Applied in The Vocational Guidance of The College Graduates. Journal of Convergence Information Technology, 8(7), pp. 876-882.

[4] Wang L (2014). Improved NN-GM(1,1) for Postgraduates' Employment Confidence Index Forecasting. Mathematical Problems in Engineering, pp. 1-8. https://doi.org/10.1155/2014/465208

[5] Liu Y, Hu L, Yan F, Zhang B (2013). Information Gain with Weight Based Decision Tree for the Employment Forecasting of Undergraduates. Green Computing and Communications, pp. 2210-2213. https://doi.org/10.1109/GreenCom-iThings-CPSCom.2013.417

[6] Kwak M, Rhee S (2016). Finding factors on employment by adult life cycle using decision tree model. 27(6), pp. 1537-1545. https://doi.org/10.7465/jkdi.2016.27.6.1537

[7] Tan L, Zhang H (2012). Forecast of Employment Based on Independent Component Analysis. International Conference on Information Computing and Applications. Springer, Berlin, Heidelberg, pp. 373-381. https://doi.org/10.1007/978-3-642-34038-3_51

[8] Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N (2019) Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomedical Signal Processing and Control, 52, pp. 456-462. https://doi.org/10.1016/j.bspc.2017.01.012

[9] Mohleji N, Mazzuchi T, Sarkani S (2014). Decision Modeling Framework to Minimize Arrival Delays from Ground Delay Programs. Air Traffic Control Quarterly, 22(4), pp. 307-325. https://doi.org/10.2514/atcq.22.4.307

[10] Zang W, Guo R (2010). Application of data mining in employment instruction of undergraduate students. IEEE International Conference on Intelligent Computing and Intelligent Systems. IEEE, pp. 772 - 775. https://doi.org/10.1109/ICICISYS.2010.5658318

[11] Nie C (2015). Forecast Analysis on Situation of College Students' Employment in China. Journal of Changchun University, 8, pp. 83-87.

[12] Qu H (2013). Influence of University Graduates Employment on Economic Growth and Its Statistical Forecast and Analysis. Journal of Applied Sciences,

13(21), pp. 4620-4623. https://doi.org/10.3923/jas.2013.4620.4623

[13] Park SH, Kim SM, Ha YG (2016). Highway traffic accident prediction using VDS big data analysis. The Journal of Supercomputing, 72(7), pp. 2832-2832. https://doi.org/10.1007/s11227-016-1624-z

[14] Li L, Zheng Y, Sun XH, Wang FS (2014). Study on Data Mining with Decision Tree Algorithm in the Student Information Management System. Applied Mechanics & Materials, pp. 3602-3605. https://doi.org/10.4028/www.scientific.net/AMM.543-547.3602

[15] Li L, Zheng Y, Sun XH, Wang FS (2016). The Application of Decision Tree Algorithm in the Employment Management System. Applied Mechanics & Materials, pp. 1252-1267. https://doi.org/10.4028/www.scientific.net/AMM.543-547.1639

[16] Hu Y (2014). The Water Conservancy Water Electricity Construction Engineering Professional Result Analysis of Application Base on C4.5 Algorithm. Advanced Materials Research, 926-930, pp. 703-707. https://doi.org/10.4028/www.scientific.net/amr.926-930.703

[17] Zhang WC, Zhang JY (2012). Vocational Graduates' Employment Problems and Countermeasures—Based on the employment of Shanxi Vocational and Technical Insistute. China University Students Career Guide, pp. 56-60.

[18] Cheng CP, Chen Q (2010). Research of Applying the Method of Decision Tree Based on Information Gain Ratio to College Student's Employment Forecasting. Computer Simulation, pp. 299-302.

[19] White T, Martin BN, Johnson JA (2003). Gender, Professional Orientation, and Student Achievement: Elements of School Culture. Journal of Women in Educational Leadership, 1, pp. 351-365.

[20] Chen CP (2007). A Research on the Employment Forecast of Graduates with Simple Bayesian Algorithm Classification. Journal of Guangdong Education Institute.

CONTENTS OF *Informatica* Volume 43 (2019) pp. 1–584

Papers

Editorials

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of S**love**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please register as an author and submit a manuscript at: http://www.informatica.si. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

SUBSCRIPTION

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than twentyfive years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica web edition is free of charge and accessible at http://www.informatica.si.
Informatica print edition is free of charge for major scientific, educational and governmental institutions. Others should subscribe.

Informatica WWW:

http://www.informatica.si/

Referees from 2008 on:

A. Abraham, S. Abraham, R. Accornero, A. Adhikari, R. Ahmad, G. Alvarez, N. Anciaux, R. Arora, I. Awan, J. Azimi, C. Badica, Z. Balogh, S. Banerjee, G. Barbier, A. Baruzzo, B. Batagelj, T. Beaubouef, N. Beaulieu, M. ter Beek, P. Bellavista, K. Bilal, S. Bishop, J. Bodlaj, M. Bohanec, D. Bolme, Z. Bonikowski, B. Bošković, M. Botta, P. Brazdil, J. Brest, J. Brichau, A. Brodnik, D. Brown, I. Bruha, M. Bruynooghe, W. Buntine, D.D. Burdescu, J. Buys, X. Cai, Y. Cai, J.C. Cano, T. Cao, J.-V. Capella-Hernández, N. Carver, M. Cavazza, R. Ceylan, A. Chebotko, I. Chekalov, J. Chen, L.-M. Cheng, G. Chiola, Y.-C. Chiou, I. Chorbev, S.R. Choudhary, S.S.M. Chow, K.R. Chowdhury, V. Christlein, W. Chu, L. Chung, M. Ciglarič, J.-N. Colin, V. Cortellessa, J. Cui, P. Cui, Z. Cui, D. Cutting, A. Cuzzocrea, V. Cvjetkovic, J. Cypryjanski, L. Čehovin, D. Čerepnalkoski, I. Čosić, G. Daniele, G. Danoy, M. Dash, S. Datt, A. Datta, M.-Y. Day, F. Debili, C.J. Debono, J. Dedič, P. Degano, A. Dekdouk, H. Demirel, B. Demoen, S. Dendamrongvit, T. Deng, A. Derezinska, J. Dezert, G. Dias, I. Dimitrovski, S. Dobrišek, Q. Dou, J. Doumen, E. Dovgan, B. Dragovich, D. Drajic, O. Drbohlav, M. Drole, J. Dujmović, O. Ebers, J. Eder, S. Elaluf-Calderwood, E. Engström, U. riza Erturk, A. Farago, C. Fei, L. Feng, Y.X. Feng, B. Filipič, I. Fister, I. Fister Jr., D. Fišer, A. Flores, V.A. Fomichov, S. Forli, A. Freitas, J. Fridrich, S. Friedman, C. Fu, X. Fu, T. Fujimoto, G. Fung, S. Gabrielli, D. Galindo, A. Gambarara, M. Gams, M. Ganzha, J. Garbajosa, R. Gennari, G. Georgeson, N. Gligorić, S. Goel, G.H. Gonnet, D.S. Goodsell, S. Gordillo, J. Gore, M. Grčar, M. Grgurović, D. Grosse, Z.-H. Guan, D. Gubiani, M. Guid, C. Guo, B. Gupta, M. Gusev, M. Hahsler, Z. Haiping, A. Hameed, C. Hamzaçebi, Q.-L. Han, H. Hanping, T. Härder, J.N. Hatzopoulos, S. Hazelhurst, K. Hempstalk, J.M.G. Hidalgo, J. Hodgson, M. Holbl, M.P. Hong, G. Howells, M. Hu, J. Hyvärinen, D. Ienco, B. Ionescu, R. Irfan, N. Jaisankar, D. Jakobović, K. Jassem, I. Jawhar, Y. Jia, T. Jin, I. Jureta, Đ. Juričić, S. K, S. Kalajdziski, Y. Kalantidis, B. Kaluža, D. Kanellopoulos, R. Kapoor, D. Karapetyan, A. Kassler, D.S. Katz, A. Kaveh, S.U. Khan, M. Khattak, V. Khomenko, E.S. Khorasani, I. Kitanovski, D. Kocev, J. Kocijan, J. Kollár, A. Kontostathis, P. Korošec, A. Koschmider, D. Košir, J. Kovač, A. Krajnc, M. Krevs, J. Krogstie, P. Krsek, M. Kubat, M. Kukar, A. Kulis, A.P.S. Kumar, H. Kwaśnicka, W.K. Lai, C.-S. Laih, K.-Y. Lam, N. Landwehr, J. Lanir, A. Lavrov, M. Layouni, G. Leban, A. Lee, Y.-C. Lee, U. Legat, A. Leonardis, G. Li, G.-Z. Li, J. Li, X. Li, X. Li, Y. Li, Y. Li, S. Lian, L. Liao, C. Lim, J.-C. Lin, H. Liu, J. Liu, P. Liu, X. Liu, X. Liu, F. Logist, S. Loskovska, H. Lu, Z. Lu, X. Luo, M. Luštrek, I.V. Lyustig, S.A. Madani, M. Mahoney, S.U.R. Malik, Y. Marinakis, D. Marinčič, J. Marques-Silva, A. Martin, D. Marwede, M. Matijašević, T. Matsui, L. McMillan, A. McPherson, A. McPherson, Z. Meng, M.C. Mihaescu, V. Milea, N. Min-Allah, E. Minisci, V. Mišić, A.-H. Mogos, P. Mohapatra, D.D. Monica, A. Montanari, A. Moroni, J. Mosegaard, M. Moškon, L. de M. Mourelle, H. Moustafa, M. Možina, M. Mrak, Y. Mu, J. Mula, D. Nagamalai, M. Di Natale, A. Navarra, P. Navrat, N. Nedjah, R. Nejabati, W. Ng, Z. Ni, E.S. Nielsen, O. Nouali, F. Novak, B. Novikov, P. Nurmi, D. Obrul, B. Oliboni, X. Pan, M. Pančur, W. Pang, G. Papa, M. Paprzycki, M. Paralič, B.-K. Park, P. Patel, T.B. Pedersen, Z. Peng, R.G. Pensa, J. Perš, D. Petcu, B. Petelin, M. Petkovšek, D. Pevec, M. Pičulin, R. Piltaver, E. Pirogova, V. Podpečan, M. Polo, V. Pomponiu, E. Popescu, D. Poshyvanyk, B. Potočnik, R.J. Povinelli, S.R.M. Prasanna, K. Pripužić, G. Puppis, H. Qian, Y. Qian, L. Qiao, C. Qin, J. Que, J.-J. Quisquater, C. Rafe, S. Rahimi, V. Rajkovič, D. Raković, J. Ramaekers, J. Ramon, R. Ravnik, Y. Reddy, W. Reimche, H. Rezankova, D. Rispoli, B. Ristevski, B. Robič, J.A. Rodriguez-Aguilar, P. Rohatgi, W. Rossak, I. Rožanc, J. Rupnik, S.B. Sadkhan, K. Saeed, M. Saeki, K.S.M. Sahari, C. Sakharwade, E. Sakkopoulos, P. Sala, M.H. Samadzadeh, J.S. Sandhu, P. Scaglioso, V. Schau, W. Schempp, J. Seberry, A. Senanayake, M. Senobari, T.C. Seong, S. Shamala, c. shi, Z. Shi, L. Shiguo, N. Shilov, Z.-E.H. Slimane, F. Smith, H. Sneed, P. Sokolowski, T. Song, A. Soppera, A. Sorniotti, M. Stajdohar, L. Stanescu, D. Strnad, X. Sun, L. Šajn, R. Šenkeřík, M.R. Šikonja, J. Šilc, I. Škrjanc, T. Štajner, B. Šter, V. Štruc, H. Takizawa, C. Talcott, N. Tomasev, D. Torkar, S. Torrente, M. Trampuš, C. Tranoris, K. Trojacanec, M. Tschierschke, F. De Turck, J. Twycross, N. Tziritas, W. Vanhoof, P. Vateekul, L.A. Vese, A. Visconti, B. Vlaovič, V. Vojisavljević, M. Vozalis, P. Vračar, V. Vranić, C.-H. Wang, H. Wang, H. Wang, H. Wang, S. Wang, X.-F. Wang, X. Wang, Y. Wang, A. Wasilewska, S. Wenzel, V. Wickramasinghe, J. Wong, S. Wrobel, K. Wrona, B. Wu, L. Xiang, Y. Xiang, D. Xiao, F. Xie, L. Xie, Z. Xing, H. Yang, X. Yang, N.Y. Yen, C. Yong-Sheng, J.J. You, G. Yu, X. Zabulis, A. Zainal, A. Zamuda, M. Zand, Z. Zhang, Z. Zhao, D. Zheng, J. Zheng, X. Zheng, Z.-H. Zhou, F. Zhuang, A. Zimmermann, M.J. Zuo, B. Zupan, M. Zuqiang, B. Žalik, J. Žižka,

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Litostrojska cesta 54, 1000 Ljubljana, Slovenia.
The subscription rate for 2019 (Volume 43) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar, borut.znidar@gmail.com.
Printing: ABO grafika d.o.o., Ob železnici 16, 1000 Ljubljana.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Slovene Society for Pattern Recognition (Vitomir Štruc)
Slovenian Artificial Intelligence Society (Sašo Džeroski)
Cognitive Science Society (Olga Markič)
Slovenian Society of Mathematicians, Physicists and Astronomers (Dragan Mihailović)
Automatic Control Society of Slovenia (Giovanni Godena)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Mark Pleško)
ACM Slovenia (Nikolaj Zimic)

Informatica is financially supported by the Slovenian research agency from the Call for co-financing of scientific periodical publications.

Informatica is surveyed by: ACM Digital Library, Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

Volume 43 Number 4 December 2019 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics

Informatica **43** (2019) Number 4, pp. 425–584