A Hybrid Approach from Ant Colony Optimization and K-nearest Neighbor for Classifying Datasets Using Selected Features

Enas M. F. El Houby, Nisreen I. R. Yassin and Shaimaa Omran
Systems & Information Department, Engineering Division, National Research Centre, Dokki, Cairo 12311, Egypt
E-mail: enas_mfahmy@yahoo.com, eng_nesrin@hotmail.com, shmomran@gmail.com

Keywords: ant colony optimization, K-nearest neighbor, features selection, heuristic, pheromone.

Received: December 12, 2016

This paper presents an Ant Colony Optimization (ACO) approach for feature selection. The challenge in the feature selection problem is the large search space that exists due to either redundant or irrelevant features which affects the classifier performance negatively. The proposed approach aims to minimize the subset of features used in classification and maximize the classification accuracy. The proposed approach uses several groups of ants; each group selects the candidate features using different criteria. The used ACO approach introduces the datasets to a fitness function that is composed of heuristic value component and pheromone value component. The heuristic information is represented with the Class-Separability (CS) value of the candidate feature. The pheromone value calculation is based on the classification accuracy resulted by adding the candidate feature. K-Nearest Neighbor is used as a classifier. The sequential forward feature selection has been applied, so it selects from the highest recommended features sequentially until the accuracy is enhanced. The proposed approach is applied on different medical datasets yielding promising results and findings. The classification accuracy is increased with the selected features for different datasets. The selected features that achieved the best accuracy for different datasets are given.

Povzetek: Opisan je hibridni pristop optimiranja s pomočjo optimizacije s kolonijami mravelj in metodo k-najbližjih sosedov.

1 Introduction

Real life data processing means having a huge amount of features that need to be analyzed, mined, classified and modeled. Classification is an important process which aims to predict the classes of future data objects. It is an automated process that requires previous knowledge of the datasets to construct a class for each group of relevant features. The aim of building classifier is to find a set of features that gives the best classification accuracy. The classification accuracy is affected by the relevancy of one feature to the other [1]. Redundant and irrelevant features worsen the performance of a classifier. This can be avoided by selecting and grouping relevant features only, thus feature selection reduces the training time and minimizes the feature set and enhances the performance of the classifier [2, 3].

The challenge in feature selection algorithm is to select minimum subset of features by eliminating features that are redundant and irrelevant which may lead the classification process to undesirable results and also removing features that do not provide predictive information. This selection is to be done with no loss of classification accuracy while reducing computation time and cost. Feature selection is an important data preprocessing phase for mining and classifying data. It is a process of selecting the optimum set of features based on a certain specified criterion to construct solid learning models [4-6]. Feature selection algorithms are divided into two categories; the first one covers the filter approach, it is an individual feature ranking approach which ranks features based on statistical methods. The second category covers the wrapper approach, which uses classifiers having classification functions to select those features with high prediction performance [7, 8].

As computation of huge number of features is not feasible; heuristic search methods are needed for feature selection. Many meta-heuristics approaches have been proposed for feature selection, such as nature inspired algorithms which have been used to select features. These algorithms like ant colony optimization have been applied to feature selection as no solid heuristic exist to find optimal feature subset, so it is expected that the ants discover good feature combinations as they proceed through the search space. Such optimization techniques were used for modeling the feature selection as an optimization problem [1, 9].

Based on this idea, in this research an Ant Colony Optimization (ACO) approach for feature selection is applied using different novel search criteria where each group of ants uses a different search criterion such as the standard deviation, the nearest, the furthest, …etc to discover different good feature combinations. In this research work, the nearest and the furthest criteria specifically were implemented. The proposed ACO approach aims to find the minimum set of features that

provide the best classification accuracy. Thus, the objective is to minimize the number of features and the classification error. The next sections are organized as follow: an overview for the previous work related to subject is presented in section 2. An introduction of ACO in feature selection problems is discussed in section 3. The proposed model is described in section 4. The experimental results are presented in section 5 before drawing conclusions and future work in section 6.

2 Related work

The feature selection and classification have become active research areas, as recently several researchers investigated various techniques to address the feature selection and classification problem. Different swarm intelligent optimizations have been used for feature selection and classification in many literatures.

Fong et al. [10] presented a study for feature selection of high dimensional biomedical datasets. They used three meta-heuristic techniques which are the particle swarm optimization (PSO), the wolf search algorithm and the bat algorithm integrated with three classification algorithms to advance the feature selection techniques and thus lowers the classification error rate. The proposed search techniques were applied on 2 biomedical datasets which are Arrhythmia and Micro Mass. Chen, et al. [11] proposed a regression based particle swarm optimization to address the feature selection problem. Regression model was used to find if the feature was selected or not using fitness values for features. Nine data sets from UCI machine learning databases were used for evaluation of the proposed algorithm. Khuat et al. [12] used directed artificial bee colony algorithm to optimize the parameters of the software effort estimation models. The accuracy of the models after optimizing parameters was improved relative to the original models accuracy. Xue et al. [13] introduced two multi-objective algorithms for feature selection based on PSO. The first one was concerned with sorting for PSO and the second one applied the crowding and mutation for PSO. They were implemented on benchmark data sets. Khazaee et al. [14] presented the PSO technique to optimize the input feature subset selection and to set the parameters for a SVM based classifier. The proposed technique was applied on three datasets for three types of electrocardiogram beats. Yeh et al. [15] presented a rule-based classifier that is constructed using simple swarm optimization, to perform the feature selection study on a thyroid gland dataset from UCI databases.

Sivagaminathan et al. [16] introduced a model that used a hybrid method of ant colony optimization and artificial neural networks (ANNs) to select features subset from medical datasets. The heuristic was calculated as a function of cost, so the feature with the lower cost was considered better and selected. Jona et al. [17] proposed a hybrid meta-heuristic search for feature selection in digital mammogram. The proposed search used a hybrid of Ant Colony Optimization (ACO) and Cuckoo Search (CS) which was used to speed local search of ACO. Support Vector Machine (SVM) was used with the ACO to classify the mammogram as normal or abnormal. Asad et al. [18] used ant colony system for features selection of retinal images dataset, a comparative study was conducted among six different features selection heuristics. They concluded that relief heuristic selection is better than the subsets selected by other heuristics. Tallon-Ballesteros et al. [19] proposed the use of Ant System (AS) search technique with two feature subset selection methods which are Correlation-based Feature Selection (CFS) and Consistency-based Feature Selection (CNS). They found that information gain is appropriate heuristic with both CFS and CNS. Dadaneh et al. [20] developed unsupervised probabilistic feature selection using ant colony optimization (UPFS). They decreased redundancy using inter-feature information which shows the similarity between the features. A matrix was used to save pheromone value between each pair of features; it was used to rank features. SVM, K-nearest neighbor (KNN), and naive Bayes (NB) were used as classifiers. Wang et al. [21] proposed a system that adjusts the parameter of ACO using different strategies to understand the parameters of ACO. The parameters included number of ants, pheromone evaporation rate, and exploration probability factor. ACO had been modified by combining it with fuzzy to be used as adaptive method for parameters. Ten UCI and StatLog benchmark data sets had been used to evaluate the performance of the proposed system. Liu et al. [22] used Bee Colony Optimization (BCO), ACO, and PSO to discover the quality of data using approaches to detect attribute outliers in datasets. The same fitness function had been used for the different search strategies. Chen et al. [23] presented an algorithm for feature selection based on ACO which traverse arcs on a directed graph; the heuristic depends on the performance of classifier and the number of selected features. SVM is used as classifier, other classifiers are used for the purpose of comparison, but SVM outperforms the other classifiers.

The introduced literatures in this section cover some of the recent researches concerned with feature selection and classification using swarm inspired algorithms. However the literatures [16-23] used specifically ACO for feature selection. Different classifiers such as ANN, SVM, KNN, and NB had been used; different heuristic such as function of cost, cuckoo search, information gain, performance of classifier and the number of selected features had been applied; different pheromone value calculation methods had been proposed; ACO parameters adjusting had been studied. And other updates had been provided to enhance the performance of ACO in features selection so as to reach the best possible accuracy with the least number of features.

The proposed work added up to these previous findings to enhance the performance of ACO in features selection. Our proposed model provides a novel idea of using different groups of ants which are synchronizing search for different solutions each using a different search criterion to reach the best possible solution for each group. Then the global best solution of all is obtained from the different applied criteria. The selection of features for different groups is done using the same fitness function but with different criteria. The fitness function is

depending on heuristic information term which is represented by Class-separability (CS) and pheromone value term which is updated using a function in the classification accuracy. The features selection is done with considering sequential forward feature selection that the feature with the best fitted fitness value is selected as it improves the classification accuracy of the selected subset; otherwise the feature with the next value is selected. A comparison between the performance of the proposed research and the previous closest work that use the same dataset will be provided in the "Experimental Results and Discussion" section.

3 Ant colony optimization

Artificial swarm intelligence is a population based meta-heuristic technique that is inspired from the behavior of living beings that live in groups, colonies, and flocks. These living organisms interact with each other and with their surrounding environment in a way that can be observed as a behavior pattern. This behavior pattern can be modeled using computers to solve different combinatorial problems. The ACO algorithm is a swarm intelligence technique that mimics real ants' attitude and behavior when searching for and grabbing their food. Ants are blind creatures that have the ability to find the shortest path from their nest to the food source. When searching for the food, ants initially explore the area surrounding their nest in a random manner. After finding the food; the ants carry part of the food and return to their nests. They release a chemical material called pheromone on the ground when moving from the food source location back to their nest. The amount of the pheromone released which mostly depends on the quantity and quality of food guides the other ants to the location of the food, and enables them to find the shortest path from their nest to the food. The ants that use a shorter path first time to grab the food returns to the nest faster releasing larger quantity of pheromone for shorter routes rather than longer routes. Afterwards, the specified shorter path will be preferred almost by all ants and as a result the pheromone starts to evaporate. The probabilistic route selection helps the ants to find the shortest routes and provide flexibility for solving optimization problems [16, 24, 25].

The ACO technique was introduced in the nineties by Dorigo et.al to solve optimization problems as the travelling salesman problem (TSP) [26]. The solution begins from a start node (feature), afterwards the ant moves iteratively from one node (feature) to the other. A probabilistic transition rule is the most widely used function in ACO which is based on the value of the heuristic function η and the pheromone value τ. It is used to iteratively construct a solution. So, the ant moves to an unvisited node (feature) with a probability of:

$$P_i^k(t) = \frac{(\tau_i(t))^{\alpha}(\eta_i(t))^{\beta}}{\Sigma_{j\epsilon N_j^k}(\tau_j(t))^{\alpha}(\eta_j(t))^{\beta}}, \quad j \in N_j^k \qquad (1)$$

Where:

N_j^k is the feasible neighborhood of the ant k, which are the features that ant k has not yet selected and can be chosen. It acts as the memory of the ants.

$\eta_i\ (t)$ is the heuristic information of the feature (i) at the time t.

$\tau_i(t)$ is the pheromone value on the feature (i) at the time t.

α and β are weights that represent the relative impact of the pheromone τ and the heuristic information η, respectively. α and β are parameters that may take real positive values according to the recommendations on parameter setting in [27].

All ants update pheromone level $\tau_i(t)$ with an increase of pheromone quantities, depending on the equations for pheromone updating, which specify how to modify the pheromone value. These equations are determined by:

$$\tau_i(t+1) = \tau_i(t) \times (1-\rho) + \Delta\tau_i(t) \qquad (2)$$

$$\Delta\tau_i^k(t) = \begin{pmatrix} Q & \text{if the feature (i) is chosen by the ant k} \\ 0 & \text{otherwise} \end{pmatrix} \qquad (3)$$

where: ρ is the pheromone evaporation rate of the features ($0 < \rho < 1$), and

$\Delta\tau_i^k$ is the pheromone deposited by the ant k that found the best solution for the current iteration.

In the ACO algorithm, ants exchange information with each other through the pheromone value. Each ant uses information obtained from other ants to provide better solutions. This improves the efficiency of solutions obtained in consecutive iterations. Thus the algorithm can be considered as a multi agent technique [28, 29].

4 The proposed ACO model

The proposed model presents an ACO approach for feature selection. Two objectives are considered; minimizing the number of features used in classification and increasing the classification accuracy. The proposed algorithm is presented in Pseudo code1.

Firstly, the algorithm starts by initializing the following parameters: The number of generations which represents the number of iterations, the number of groups of ants which represent the number of different criteria for features selection, the number of ants which is equivalent to the number of solutions for each criteria (group), maximum number of features that represent maximum allowed number of features that can be selected by each ant to achieve the best possible classification accuracy. (τ_i) which is the pheromone concentration value associated with feature (f_i) and (η_i) is the heuristic value for feature (f_i); (τ) & (η) together form the fitness function terms as shown in eq. (1). α, β are user selected parameters; they represent the relative importance of pheromone and heuristic values respectively. ρ is a user defined parameter that represents the pheromone evaporation or decay rate, it takes a value from 0 to 1. Z is the local pheromone update parameter, it is defined by the user and it takes a value less than ρ.

A set of generations starts after the initialization phase. With each new generation, n groups of ants are formed where G_1, G_2,….G_n are n different groups each having na ants. The first feature selection for each ant is

performed randomly taking into consideration avoiding redundancy between ants of the same group to obtain different possible solution sets. For each selected feature by different ants an initial value for the classification accuracy is obtained for each ant using the KNN algorithm. Using a set of equally initial pheromone values, and the local pheromone update parameter Z, the pheromone of the selected feature is locally updated using eq. (4).

$$\tau_i(t+1) = (1-Z) \times \tau_i(t) + Z \quad (4)$$

For each group of ants, the selection of the subsequent features is done using the same fitness function with different criteria. One used criterion is to select the nearest feature to the previously selected one according to fitness value. Another used criterion is to get the furthest feature to the previously selected one according to fitness value. The fitness function is calculated using eq. (1) having a term representing the pheromone τ and a heuristic term η. The heuristic information is represented with the Class-Separability (CS) value of the feature. All features have equally initial pheromone values and the pheromone of the selected feature is locally updated with eq. (4). By the end of each generation, the pheromone values of the features subsets that are part of the best solution for different groups are globally updated using eq. (5). It is a function in the classification accuracy achieved by selected features subset, so as to increase the features selection opportunity of these features in the future generations.

$$\tau_i(t+1) = (1-\rho) \times \tau_i(t) + \rho * acc \quad (5)$$

Where: ρ is the pheromone evaporation rate of the features ($0 < \rho < 1$), and acc is the classification accuracy.

As mentioned above, the selection of the subsequent feature is done using different criteria. This selection is performed using sequential forward selection. The next fitted feature is selected if it improves the classification accuracy and it is considered positively correlated with the preceding features. Otherwise, if the feature reduces the accuracy or maintains it constant, it is considered negatively correlated or neutral and is not selected. This selection is repeated until finding the feature that satisfy the group criteria whether nearest or furthest and improve the classification accuracy. This process is repeated for selecting each subsequent feature. The stopping criteria is either obtaining the subset that achieves the best possible accuracy or reaching the maximum allowed number of features for the different ants. By the end of generation, the pheromone values of the features that are part of the best solution are updated.

After that, a new generation is started with the updated pheromone values for features to generate different features subsets in the next generation. By the end of all generations, the features subsets that give the best accuracy in all generations and by different ants for each group are obtained, and then the best global subsets by different groups of ants are obtained. Figure 1 illustrates the full process of feature selection using the proposed ACO model.

4.1 Class-separability

As mentioned previously, the heuristic value is computed using the class-separability approach. Class-separability (CS) [30] is an approach used for feature selection. CS of feature i is defined as:

$$CS_i = SB_i / SW_i \quad (6)$$

Where

$$SB_i = \sum_{k=1}^{K} (\bar{x}_{ik} - \bar{x}_i)^2 \quad (7)$$

$$SW_i = \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 \quad (8)$$

$$\bar{x}_{ik} = \sum_{j \in C_k} x_{ij} / n_k \quad (9)$$

$$\bar{x}_i = \sum_{j=1}^{n} x_{ij} / n \quad (10)$$

SB_i is the sum of squares of between class distances (the distances between samples of different classes). SW_i is the sum of squares of within class distances (the distances of samples within the same class). In the whole data set, there are K classes. C_k refers to class k that includes n_k samples. x_{ij} is the value of feature i in sample j. $\bar{x}_{ik}$ is the mean value in class k for feature i. n is the total number of samples. $\bar{x}_i$ is the general mean value for feature i. A CS is calculated for each feature. A larger CS indicates a larger ratio of the distances between different classes to the distances within one specific class. Therefore, CS can be used to measure the capability of features to separate different classes.

4.2 K-Nearest Neighbor

KNN (also known as Instance-based Learning) is a simple efficient data mining technique used for classifying data points based on their distance to other points in a training dataset. KNN is a lazy learner where the training data is loaded into the model and when a new instance need to be classified it looks for the specified k number of nearest neighbors; then, takes a vote to see where, the instance should be classified. For example, if k is 7, then the classes of 7 nearest neighbors are detected. KNN depends on a simple principle which is "similar instances have similar class labels". Distance functions are used to measure similarity between samples. KNN calculates the distance between the unknown data point and every known data point [31, 32]. The common distance between two data points $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_n)$ is defined as follows:

$$Dist(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \quad (11)$$

Euclidian distance is the easiest and most common distance calculation function for quantitative data where $p = 2$.

```
Initialize parameters [number of generations, number of groups of ants, number of ants,
max number of features, pheromone value (τ), heuristic value(η), pheromone evaporation
rate(ρ),local pheromone update parameter(Z), α, β]

While Generation number not exceeded
   For each group of ants
      For each ant
         o Select a start distinct feature randomly
         o Calculate the classification accuracy for the initially randomly selected
           feature
      End for
   End for

   For each ant in each group of ants
      While assigned number of features not exceeded (max number of features) and
      accuracy can still be improved
         o Calculate the fitness value for each feature that can be selected using
           eq.(1)
         o Each group of ants select the candidate feature using different criteria
         o Calculate the accuracy for the subset obtained with the new added feature
         o If the calculated accuracy less than or equal the previous accuracy select
           consecutive feature sequentially with next appropriate fitness value
         o If the maximum accuracy is achieved or max number of feature is reached;
           Final set of features is obtained for the current ant
           Else update the pheromone value of the selected feature locally by eq.(4)
           End If
      End while
   End for

   o Find the features subsets that achieved the best solution in the current generation
     for different groups
   o Update the pheromone values of the features that are part of the best solution
     using eq.(5)to increase these features selection opportunity in next generations
End while
Obtain the global best collected subsets of features from all different groups of ants
that achieved best possible accuracy in all generations
```

Pseudo code 1: The proposed ant colony algorithm for feature selection.

5 Experimental results and discussion

This section shows an empirical performance evaluation of the proposed ACO model. In order to evaluate the proposed model, real world medium-sized datasets shown in Table 1 are tested. The used datasets are heart disease, breast cancer and thyroid which contain 13, 30, 21 features and 303, 569, 7200 samples respectively. The samples are randomly divided into 257/46, 423/146 and 6000/1200 for training and testing respectively. Matlab® 2015a software on an Intel®Core™ i7 CPU @ 1.6 GHz computers is used for implementation. Extensive experimental studies had been tried in order to get the best features subsets selected by ACO which give the highest possible accuracy using KNN classifier. The values of the parameters have been tuned in the experiments as shown in Table 2.

Different features subsets have been tried starting from 2 features subsets until the maximum number of features specified by the user is reached or the best possible accuracy is achieved. In each trial, two different groups of ants, each consists of 3 ants, start to select the features. A number of 100 iterations or generations are executed to reach the highest possible accuracy. Each of the two groups of ants uses a different criterion to select the features. The first group uses the nearest feature to the previously selected one according to the fitness value. The other group uses the furthest feature to the previously selected one according to the fitness value. The pheromone of the selected feature has been increased by small value using eq. (4). At the end of each iteration, the pheromone value of the features that are part of the best solution of either group is incremented by a significant value calculated using eq. (5).

Table 3 illustrates the idea of the proposed ACO model for features selection of breast cancer dataset using two ant groups; nearest and furthest. For illustration purpose, the table includes samples of features subsets with the accuracy which had been recorded as the best accuracy in different generations to clarify the idea. For example, if by the end of a generation, the best achieved accuracy was 93.84% using the two groups of ants. The nearest group achieved it using the features subset {8, 29} and the furthest group achieved it using {24, 4} subset, so the pheromone value of both group subsets will be updated for that generation. The pheromone value of these features will be increased using eq. (5) to have a chance to be part of the best solution in their groups in the next generation.

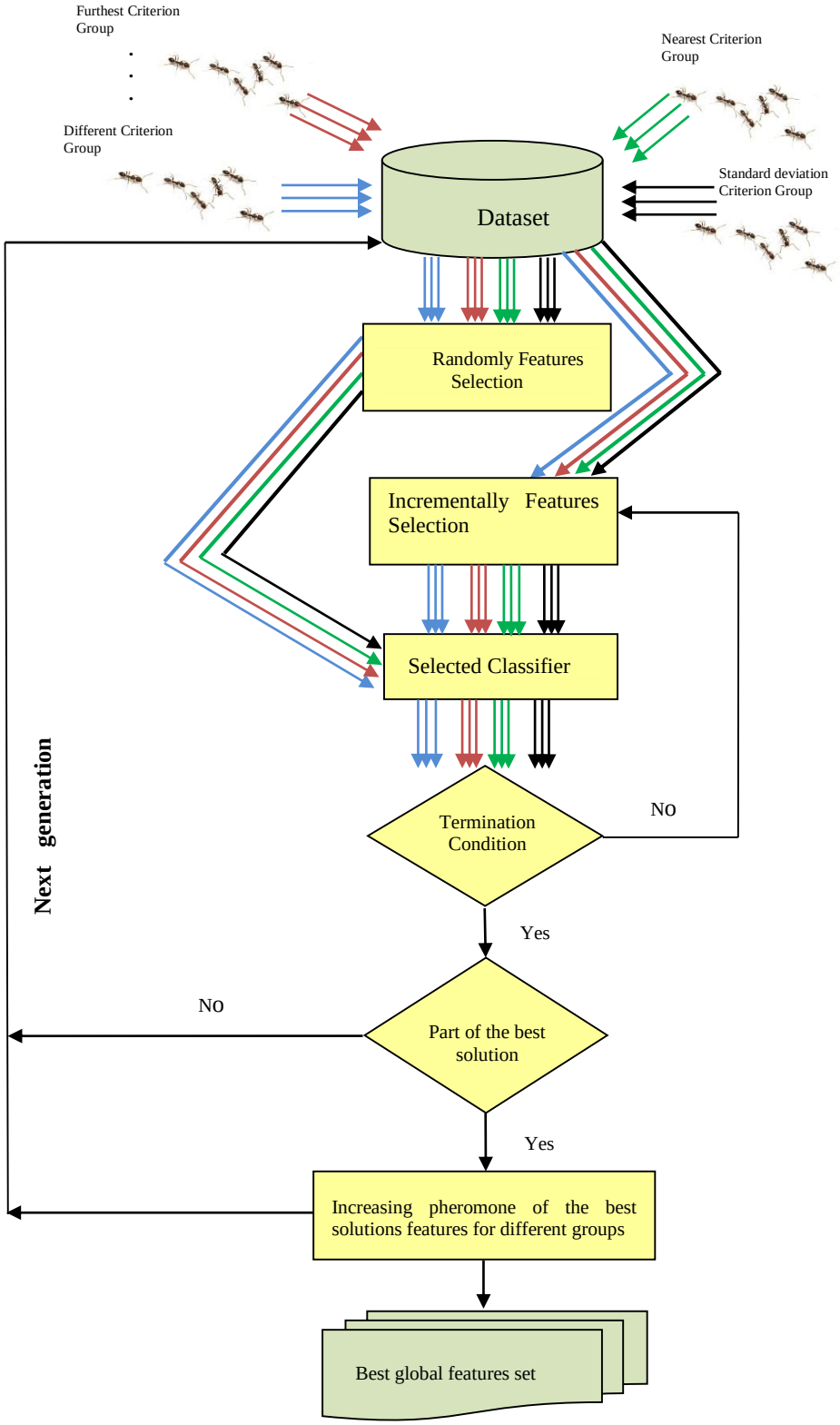

Figure 1: The full process of feature selection using the proposed ACO approach.

Table 1: The used datasets.

Data set name	No. of features	No. of samples	No. of classes	Citation
Heart Disease	13	303	5	[33]
Breast cancer (Wisconsin diagnostic)	30	569	2	[34]
Thyroid	21	7200	3	[35]

In many cases, the same subset can be reached by the two groups either in the same generation or through different generations; this may be due to the limitation of selecting positively correlated features. Also the nearest or the furthest features may change with the generations. This is due to the local changes of the pheromone's values of the selected feature and global changes of the best features by the end of generation which causes the changing in the distances between features. As an example, the 2 subsets that achieved 94.52 % accuracy which are {8,29,30; 8,29,6}. In the first set the furthest feature to the feature 29 was 30 and in the second set it was 6.

Table 2: Parameters of ACO.

Parameters	Values
α	1
β	0.9
ρ	0.9
Z	0.4
No. of generations	100
groups of ants (G)	2
no. of ants/G (na)	3

The best global features subset in the breast cancer dataset that achieved the best possible accuracy which is 97.95% with the least number of features which is 4 features is {16, 28, 7, 1}. By increasing the number of features above 4, the accuracy remains constant as shown in Table 4. So, the selected 4 features are considered the best combination of features that satisfy the best possible accuracy. Figure 2 present the relation between the accuracy versus number of features for breast cancer dataset. The selected features that achieved the best possible accuracy which is 97.95% are the features of cell nucleus mentioned below, these features take decimal values:

1 - mean radius (mean of distances from center to points on the perimeter)

7 - mean of concavity (severity of concave portions of the contour)

16 - standard error of compactness (perimeter^2/area-1.0)

28 - worst concave points (number of concave portions of the contour)

Regarding the heart disease dataset, as shown from Table 5, although the best possible accuracy which is 96.88% has been achieved by 4 features, the best possible accuracy using 2 features is close to it which is 96.77% and it needs only 2 features. So, the set {9, 12} is recommended as best solution. With increasing the features set above 4 features, the accuracy decreases as shown in Table 5. The best selected features are "number of major vessels colored by fluoroscopy" and "exercise induced angina". The first feature takes a value from 0 to 3, the second feature is Boolean and takes values (1 = yes; 0 = no). Figure 3 shows the relation between the accuracy versus number of features for heart disease dataset.

For the thyroid dataset, as shown in Table 6, although the best possible accuracy which is 98.5% has been achieved by 6 features, the best possible accuracy using 4 & 5 features is close to it which is 98.25%, 98.33% respectively. With increasing the features set to 7 features the accuracy remains constant. Figure 4 shows the relation between the accuracy versus number of features for thyroid dataset. The best selected features are TSH (real [0.0 - 0.53]), thyroid surgery (integer [0, 1]), on thyroxin (integer [0, 1]), and FTI (real [0.0020 - 0.642]).

Table 7 shows the percentage of features reduction and the achieved accuracy before and after features reduction for different datasets. Figure 5 presents the total number of features and the reduced number of features using the proposed ACO model for different datasets. Figure 6 presents the comparison of the accuracy for the total number of features and the reduced number of features for the three datasets used. It is clearly that the selected features achieve higher accuracy than the total number of features which ensure that noisy and irrelevant features mislead the classifiers.

Table 8 shows the comparison of the proposed model with the previous work. Since the main purpose of features selection is to achieve the highest accuracy with the least number of features, a comparison between the performance of the proposed research and the previous closest work, will be limited to those that used the same dataset to simplify the comparison capability. By comparing the proposed model with previous work, it seems that the proposed model outperforms others with even less number of features for all databases. Except for breast cancer, Wang G., et al. [21] achieved 98.12% accuracy which is a bit better than ours (97.95%), but with larger numbers of features which are13.5 features rather than only 4 features with our suggested algorithm.

To investigate the capability of the proposed model to achieve promising results with large dataset, the SRBCT microarray dataset [36] which contains 2308 features (genes) and 83 samples was used. The samples are divided into 63/20 for training and testing respectively. It achieved 100% accuracy with 4 genes only as shown in Table 9, the percentage of features reduction is 99.82%. After applying the proposed model on SRBCT dataset, it is concluded that it also has the ability to select features subset which

Table 3: samples of selected features subsets and achieved accuracy using ACO model through different generations for breast cancer dataset.

No. of feature	nearest	Best Acc. %	furthest	Best Acc. %
2	{8,29; 24,4}	93.84	{24,4; 8,29; 6,16; 23,1}	93.84
2	{18,28; 1,21}	94.52	{2,23; 18,28; 21,1}	94.52
2	{4,23; 26,1}	95.21	{4,23}	95.21
3	{ 27,3,23; 8,29,30; 8,29,9; 26,21,3}	94.52	{8,29,30; 8,29,6; 20,23,2; 1,23,12; 18,23,2;16,23,2}	94.52
3	{ 28,18,30; 23,3,14; 21,12,1}	95.89	{14,3,23; 3,23,14}	95.89
3	{ 1,28,7}	97.26	{7,28,1}	97.26
4	{12,21,3,1; 3,23,14,2; 1,21,29,12}	96.58	{ 21,12,1,13; 8,29,19,10; 12,20,27,3}	96.58
4	{7,28,1,16}	97.95	{16,28,7,1}	97.95
5	{28,8,7,1,17; 5,7,28,1,17}	97.26	-	97.26
5	-	97.95	{10,28,7,1,16; 30,28,7,1,16}	97.95
6	{18,7,16,20,28,1;28,16,7,18,10,1}	97.95	{19,28,7,1,17,30}	97.95
7	{25,7,28,11,17,16,10;18,29,6,21,1,12,27}	97.26	-	-
7	-	-	{19,28,10,7,17,1,30}	97.95

Table 4: The selected features subsets using the proposed ACO model that achieved the best accuracy/feature subset for breast cancer.

No. of selected features	**The reduced features subsets**	**Best Accuracy (%)**
2	{4,23; 26,1}	95.21
3	{7,28,1}	97.26
4	{16,28,7,1}	97.95
5	{10,28,7,1,16; 30,28,7,1,16; 15,28,7,1,16; 7,28,30,1,17}	97.95
6	{18,7,16,20,28,1; 28,16,7,18,10,1; 19,28,7,1,17,30}	97.95
7	{19,28,10,7,17,1,30}	97.95
Most important features subset	Features of cell nucleus: 1 - mean radius 7 - mean of concavity 16 - standard error of compactness 28 - worst concave points	

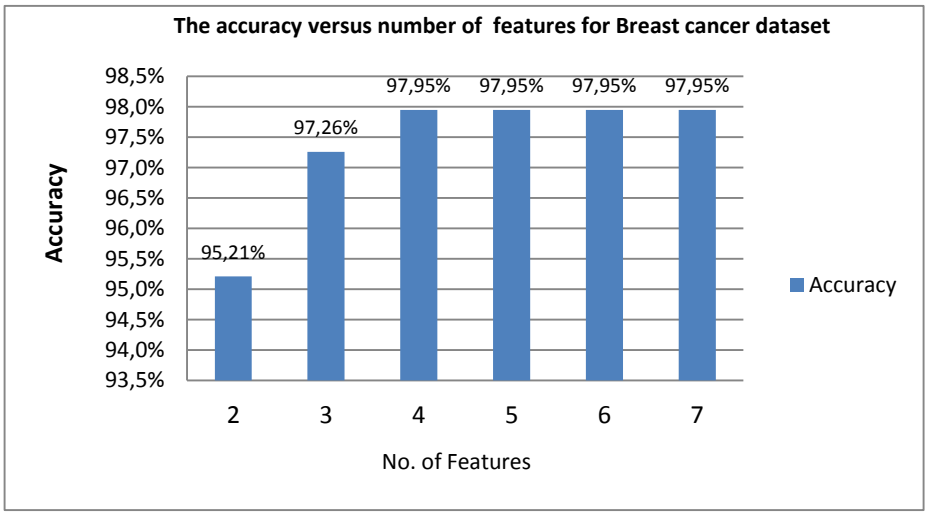

Figure 2: The accuracy versus number of features for breast cancer dataset.

achieve the highest accuracy from large number of features.

Computational Complexity

The actual computational cost of an algorithm can be determined by investigating the computational complexity according to the form of big-O notation. Meta-heuristic algorithms are simple in terms of complexity, and thus they are easy to implement. ACO algorithm has three inner loops n the number of groups of ants; na is the number of ants; and f is the number of selected features, and one outer loop t for iteration. So, the time complexity is $O(n * na * f * t)$. In the experimental studies the inner loops are small (n = 2; na =3; F = 2-7) and (t =100), so the computational cost is relatively inexpensive. The main computational cost will be in five steps according to Pseudo code1: (i)

Table 5: The selected features subsets using the proposed ACO model that achieved the best accuracy /feature subset for heart disease.

No. of selected features	The reduced features subsets	Best Accuracy (%)
2	{9,12}	96.77
3	{6,12,9; 4,12,1}	93.75
4	{4,12,1,6}	96.88
5	{4,12,6,10,2}	94.12
6	{4,6,5,7,1,12}	93.1
7	{4,6,5,7,3,8,12; 4,13,2,10,3,8,12}	84.62
Most important features subset	12 - ca: number of major vessels (0-3) colored by fluoroscopy 9 - exang: exercise induced angina (1 = yes; 0 = no)	

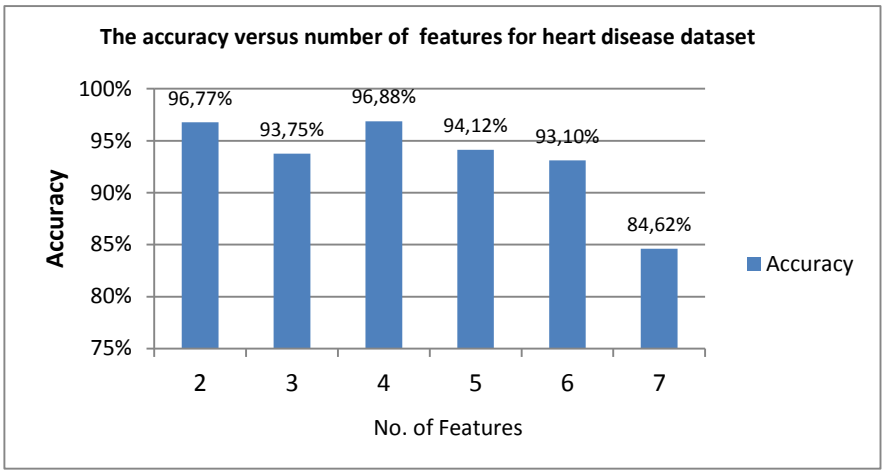

Figure 3: The accuracy versus number of features for heart disease dataset.

Table 6: The selected features subsets using the proposed ACO model that achieved the best accuracy/feature subset for Thyroid.

No. of selected features	The reduced features subsets	Best Accuracy (%)
2	{3,17}	97
3	{21,17,3}	97.92
4	{8,17,3,21}	98.25
5	{19,17,5,3,8; 8,17,3,21,16; 3,17,6,21,8}	98.33
6	{21,17,9,3,16,8}	98.5
7	{3,17,12,21,9,16,8; 21,17,5,3,9,8,16}	98.5
Most important features subset	17 - TSH 8 - Thyroid_surgery 3 - On_thyroxine 21 - FTI	

random feature selection, (ii) subset selection using different criteria, (iii) updating pheromone values (iv) calculating probabilistic transition rule, and (v) termination condition.

For example, the estimated time for heart disease to select 2 features was ≅ 15.3 sec on average, selecting 3 features needs 35.5 sec on average and selecting 4 features needs 48 sec on average.

6 Conclusion and future work

In this research, an ACO model has been developed to select minimum subsets of features from datasets that can achieve the best possible accuracy. The purpose was to reduce redundant, irrelevant and noisy features that mislead the classifier. The proposed model use different groups of ants to select different features subsets that give the best possible result. Each group uses a different criterion to select the features. In this research two different criteria have been applied; the nearest and furthest criteria. By the end of each generation, each group selects the best features subsets that achieve the best accuracy for that group. The pheromone values of these features are increased to be given higher opportunity to be part of the selected features in the next generation for that group. By the end of all generations, the best features subsets for each group have been selected, and then the global best solutions from all groups have been reached.

The results showed that, the right selection of features and eliminating irrelevant, noisy and redundant features increase the accuracy of classifiers. The percentage of

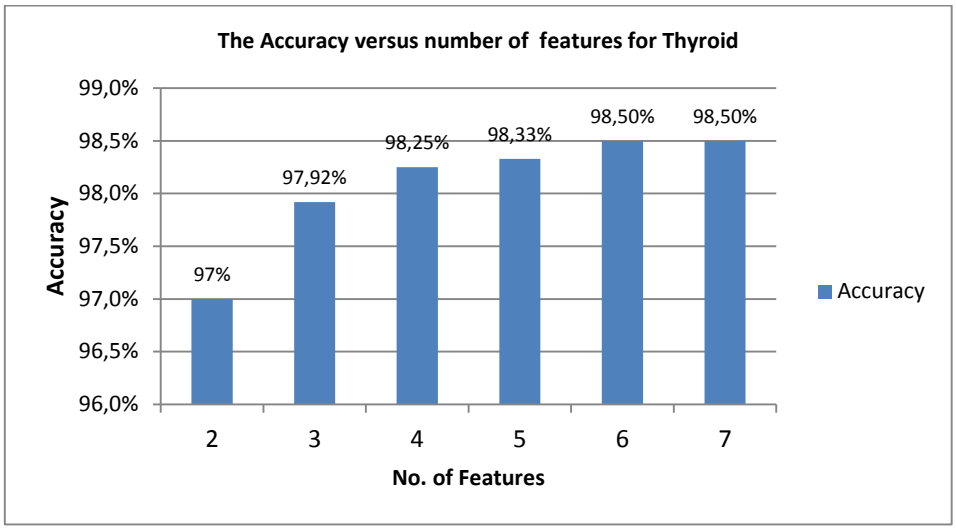

Figure 4: The accuracy versus number of features for Thyroid dataset.

Table 7: The percentage of features reduction and the achieved accuracy for different datasets.

Data sets	Training size/Testing	No. of features	Reduced subset	% Reduction	Accuracy of total features	Accuracy of reduced features
Heart disease	257/46	13	2	84.61 %	82.5%	96.77%
Breast cancer (Wisconsin diagnostic)	423/146	30	4	86.66 %	93.15 %	97.95 %
Thyroid	6000/1200	21	4	80.95 %	93.15%	98.25 %

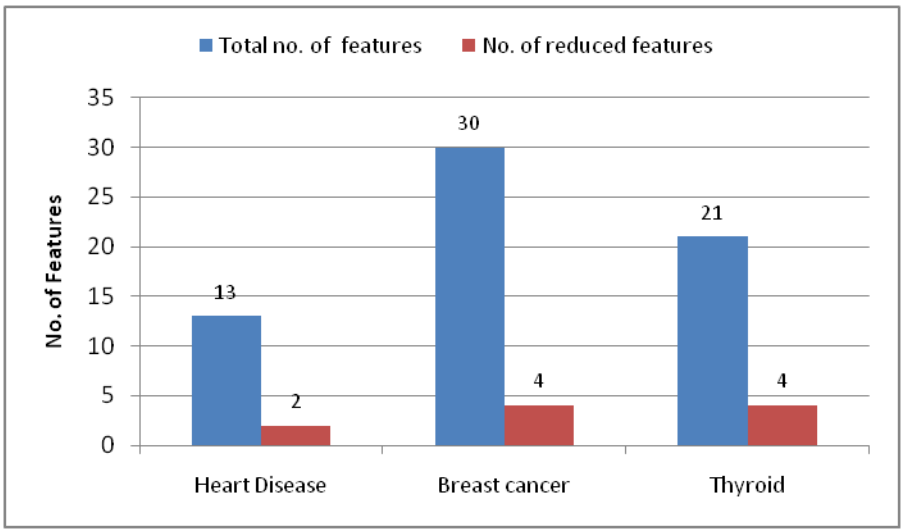

Figure 5: The total no. of features compared with reduced no. of features for different datasets.

features reductions are 84.61 %, 86.66 %, and 80.95 % for heart disease, breast cancer, and thyroid respectively with that the accuracy had increased from 82.5% to 96.77%, from 93.15 % to 97.95 % and from 93.15% to 98.25 % respectively. By trying the model on different datasets it achieved promising results compared with previous works, it achieved higher accuracy with less number of features for all databases. Different features subsets have been reached using different groups' criteria which give the capability to collect different solutions and reach the best global solutions. The proposed model proved its capability to select features from large datasets, when applied on SRBCT microarray dataset. As a future work, other criteria can be used to collect different subsets of features, also parametric studies can be studied and applying the model on different datasets can be done.

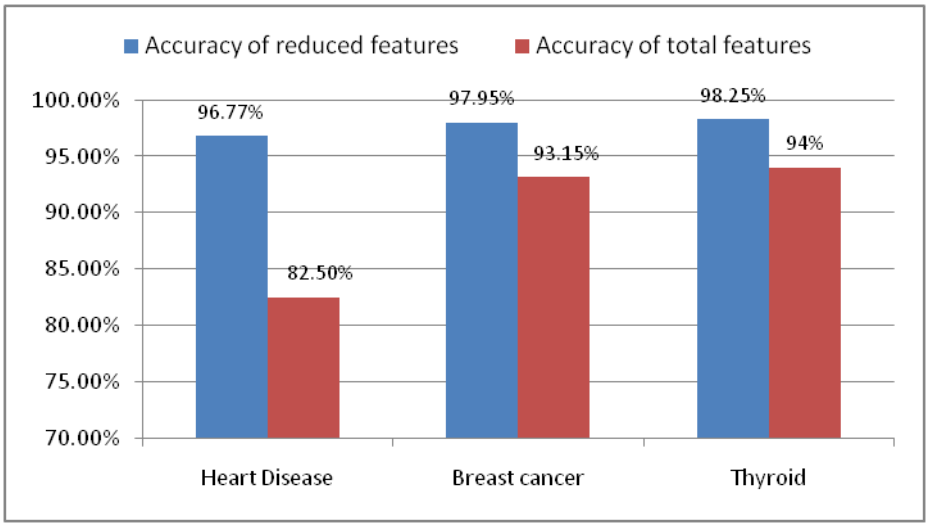

Figure 6: The comparison of the accuracy for reduced no. of features and the total no. of features.

Table 8: The comparison of the proposed model with the previous work.

Data sets	No. of features	No. of Reduced Features/ Accuracy				
		Proposed system	Sivagaminathan, et al. [16]	Dadaneh, et al. [20]	Wang G., et al. [21]	Chen, et al. [23]
Heart Disease	13	2 / 96.77 %	-	-	6.1/88.22% On average	8.08/ 86.67% On average
Breast cancer (Wisconsin diagnostic)	30	4 / 97.95 %	12 / 95.57 %	11 / 91.23%	13.5/98.12% On average	5.89 / 95.99 On average
Thyroid	21	4 / 98.25 %	14 / 94.5 %	-	-	-

Table 9: The selected features subsets using the proposed ACO model that achieved the best accuracy /feature subset for SRBCT.

No. of selected features	The reduced features subsets	Best Accuracy (%)
2	{1613, 1165; 1427, 1479}	85
3	{156, 1606, 1601}	95
4	{1434, 1775, 2214, 1920}	100
Most important features subset	1434- kinesin family member 3C 1775- guanosine monophosphate reductase 2214- FYN oncogene related to SRC, FGR, YES 1920- lamin B2	

7 References

[1] S. M. Vieira, J. M. Sousa, and T. A. Runkler, (2009). Multi-criteria ant feature selection using fuzzy classifiers. In Swarm Intelligence for Multi-objective Problems in Data Mining, Springer Berlin Heidelberg, 19-36.

[2] I. A. Gheyas, L. S. Smith, (2010). Feature subset selection in large dimensionality domains. Pattern Recognition 43(1) 5-13.

[3] A. Unler, A. Murat, (2010). A discrete particle swarm optimization method for feature selection in binary classification problems. Eur. J. Oper. Res. 206(3) 528–539

[4] M. Dash, K. Choi, P. Scheuermann, and H. Liu, (2002). Feature selection for clustering filter solution. In: Proc. of Second International Conference on Data Mining ICDM 115–122.

[5] P. Mitra, C. A. Murthy, and S. K. Pal, (2002). Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Analysis and Machine Intelligence 24(3) 301–312

[6] A. Miller, (2002). Subset Selection in Regression. (2nd ed.). Chapman & Hall/CRC, Boca Raton.

[7] Blum, L. Avrim and P. Langley, (1997). Selection of relevant features and examples in machine learning. Artificial intelligence, 1, 245-271.

[8] L. Talavera, (2005). An evaluation of filter and wrapper methods for feature selection in categorical clustering. In International Symposium on Intelligent Data Analysis, Springer Berlin Heidelberg, 440-451.

[9] L. A. M. Pereira, D. Rodrigues, T. N. S. Almeida, C. C. O. Ramos, A. N. Souza, X-S. Yang, and J. P. Papa, (2014). A Binary Cuckoo Search and Its Application for Feature Selection, In Cuckoo Search and Firefly Algorithm. Springer International Publishing 141-154.

[10] S. Fong, S. Deb, X. S. Yang, and J. Li, (2014). Feature selection in life science classification: metaheuristic swarm search. IT Professional 16(4) 24-29.

[11] K. H. Chen, L. F. Chen, and C. T. Su, (2014). A new particle swarm feature selection method for classification. Journal of Intelligent Information Systems 42(3) 507-530.

[12] Thanh Tung Khuat, My Hanh Le, (2016).Optimizing Parameters of Software Effort Estimation Models using Directed Artificial Bee Colony Algorithm, Informatica 40, 427–436.

[13] B. Xue, M. Zhang, and W. N. Browne, (2013). Particle swarm optimization for feature selection in classification: a multi-objective approach, IEEE transactions on cybernetics 43(6) 1656-1671.

[14] A. Khazaee, (2013). Heart beat classification using particle swarm optimization, International Journal of Intelligent Systems and Applications 5(6) 25.

[15] W. C. Yeh, Novel swarm optimization for mining classification rules on thyroid gland data, Information Sciences 197 (2012) 65-76.

[16] R. K. Sivagaminathan, and S. Ramakrishnan, (2007). A hybrid approach for feature subset selection using neural networks and ant colony optimization. Expert systems with applications 33(1), 49-60.

[17] J. B. Jona, and N. Nagaveni, (2014). Ant-cuckoo colony optimization for feature selection in digital mammogram. Pakistan Journal of Biological Sciences 17(2), 266.

[18] A. Asad, A. T. Azar, N. El-Bendary, and A. E. Hassaanien, (2014). Ant colony based feature selection heuristics for retinal vessel segmentation. arXiv preprint arXiv:1403.1735.

[19] A. J. Tallon-Ballesteros and J. C. Riquelme, (2014). Tackling Ant Colony Optimization Meta-Heuristic as Search Method in Feature Subset Selection Based on Correlation or Consistency Measures. in International Conference on Intelligent Data Engineering and Automated Learning. 2014. Springer, 386–393.

[20] Behrouz Zamani Dadaneh, Hossein Yeganeh Markid, Ali Zakerolhosseini, (2016). Unsupervised probabilistic feature selection using ant colony optimization. Expert Systems with Applications, 5327–42.

[21] G. Wang, H. E. Chu, Y. Zhang, H. Chen, W.Hu, Y. Li, and X. J. Peng, (2015). Multiple parameter control for ant colony optimization applied to feature selection problem. Neural Computing and Applications 26(7), 1693-1708.

[22] Bo Liu, Mei Cai and Jiazong Yu, (2015). Swarm Intelligence and its Application in Abnormal Data Detection. Informatica 39(1), 63–69.

[23] B. Chen, L. Chen, and Y. Chen, (2013). Efficient ant colony optimization for image feature selection. Signal processing 93(6) 1566-1576.

[24] C. Coello, S. Dehuri, and S. Ghosh, eds, (2009). Swarm intelligence for multi-objective problems in data mining. 242 Springer.

[25] Kanan, H. Rashidy, K. Faez, and S. M. Taheri, (2007). Feature selection using ant colony optimization (ACO): a new method and comparative study in the application of face recognition system. In Industrial Conference on Data Mining, Springer Berlin Heidelberg, 63-76.

[26] M. Dorigo, (1992). Optimization, learning and natural algorithms, Ph.D. Thesis, Politecnico di Milano, Italy.

[27] M. Dorigo, V. Maniezzo, A. Colorni, A., (1996). Ant system: optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 26(1), 29-41.

[28] S. Tabakhi, M. Parham, and F. Akhlaghian, (2014). An unsupervised feature selection algorithm based on ant colony optimization. Engineering Applications of Artificial Intelligence 32, 112-123.

[29] S. Dehuri, S. Ghosh, and C. A. Coello, (2009). An introduction to swarm intelligence for multi-objective problems. In Swarm Intelligence for Multi-objective Problems in Data Mining, Springer Berlin Heidelberg, 1-17.

[30] Feng chu & lipo wang, (2005). Applications of support vector machines to cancer classification with microarray data. International Journal of Neural Systems, 15(6), 475–484

[31] C. C. Aggarwal, 2015. Data Mining: The Textbook, Springer International Publishing Switzerland.

[32] M. Kubat, 2015. An Introduction to Machine Learning, Springer International Publishing Switzerland.

[33] http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[34] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[35] http://sci2s.ugr.es/keel/category.php?cat=clas

[36] Alexander Statnikov, C.F.A., Ioannis Tsamardinos. Gene Expression Model Selector. 2005 [cited 2017 April]; Available from: www.gems-system.org