Attribute Induction-oriented Excavation and Generalization Analysis of Site Archaeological Data

Jianbing Zhang
Mechanical and Electrical Engineering Department, Zhumadian Vocational and Technical College, Zhumadian 463000, China
E-mail: zjb19780815@163.com

Keywords: Site archaeology; Data excavation; Generalisation analysis; K-means; Apriori; Attribute induction

Received: March 14, 2023

A significant amount of digital archaeological data has emerged as a result of the recent increase in archaeological activity, which is crucial for the preservation of cultural heritage. However, redundant and repetitive archaeological data information often leads to difficulties in management. The study first enhances the Apriori algorithm, which is based on determining the artefact attributes of site archaeological data by applying the boosting degree with difference, in order to increase the effectiveness of archaeological research. A K-means algorithm with adaptive selection of initial clustering centres was then proposed as a means to generalise the archaeological data for analysis. The outcomes revealed that the enhanced Apriori algorithm's maximum runtime was only 0.33 seconds and its minimum runtime was 0.1 seconds. Due to the low impact of noise points on the dataset Flame, the revised K-means algorithm's standard deviation is only 2.537, with the majority of the error values being clustered around zero. After combining the two methods, the classification accuracy of the digitised resources of the site is concentrated around 92%, with high classification accuracy and data generalisation processing ability, which improves the processing efficiency and provides a more reliable method reference for site archaeological research efficiency improvement.

Povzetek: Študija uporablja izboljšan Apriori algoritem za določanje atributov artefaktov in prilagojen K-means algoritem za generalizacijo arheoloških podatkov.

1 Introduction

Human society has left behind a vast amount of material materials in production and life, and with the changes of the times, these valuable heritages have become the witness of social development, providing an important basis for human beings to realise their cultural heritage [1,2]. In the process of scientific investigation and excavation, archaeology is an important driving force in historical research by systematically and completely revealing and collecting the relics buried deep in the ground and uncovering the historical and cultural values and artistic values they contain. Digital archaeology is continuing to progress, and digital resources are expanding rapidly due to the research and technological advances that are occurring so quickly. These digital resources are used throughout the archaeological excavation process, and not only have a huge amount of data, but also a great variety, which needs to be managed and utilised effectively. At the same time, the majority of archaeologists are still in the manual management and searching stage and are unable to keep up with the demand for the enormous amount of archaeological data [3,4]. How to integrate various archaeological data sources to boost archaeology's efficacy is a significant challenge for archaeologists in the information era. Data mining technology is capable of uncovering latent patterns, and its integration of database, machine learning and other multi-disciplinary techniques is one of the effective methods for conducting massive random data mining nowadays. Data generalisation for site archaeology can help archaeologists to view data at different levels in a customised degree of abstraction, greatly improving the efficiency of archaeological data analysis [5,6]. Applying data mining techniques to archaeological data generalisation analysis has a high degree of feasibility. To further increase the effectiveness of archaeological labor, the study suggests using the Apriori algorithm and K-means algorithm to generalize archaeological data by identifying the characteristics of site artifacts.

2 Literature review

Archaeologists have recently paid a lot of attention to the excavation of archaeological data from sites. Previtali and Valente developed a framework for sharing archaeological data in order to maximise the impact of archaeological data, which utilises digital technology for the collection of raw data and realises the processing of raw data through image processing and scan alignment, and the results show that the method enriches the raw assemblage data with a high degree of interoperability

[7].

Korf et al. developed a non-targeted data analysis method based on mass spectrometric detection of compounds in archaeological samples, which extracted more compounds through isotopic analysis of residual molecules in the samples, and showed that the method reduced the researcher's bias against extracting too few compounds and accelerated the overall analysis time [8]. Tronicke et al. developed a multi-scale, analysis and visualisation method to address archaeological practice and research. This method produces a discrete redundant wavelet transform (RWT) using a à trous algorithm and the cubic spline filter, then computes a multiscale decomposition of 2D data via a sequence of 1D convolutions. The results demonstrate that the approach simplifies the understanding of archaeological geophysical datasets and is computationally efficient [9]. Kansa and Kansa address the issue of digital data in archiving, modelling and other issues in archaeological practice, proposes a data management approach that incorporates multiple techniques and explores the creation of archaeological data for a wider range of needs, providing a new reference for the creation of archaeological information systems [10]. Otárola-Castillo et al. addresses archaeologists' frequent excavation and investigation of archaeological data through null hypothesis significance tests and probability distributions, proposing a method for accomplishing this using Bayesian statistics. It creates a Bayesian statistical framework for handling archaeological data and employs a collection of data for a particular site, and the findings show the method's viability in a real-world setting [11]

The improvement of clustering algorithms has some reference value for site archaeological data excavation processing. In response to the sensitivity of the initial cluster center selection of the K-means algorithm, Li's team proposed an improved hybrid particle swarm optimisation algorithm for clustering centers. The outcomes demonstrated that the approach had a high accuracy rate and enhanced convergence speed [12,13]. Using the algorithm's pseudo-code and experimental validation on a standard dataset (Lris), to be sure that the K-value selection had no impact on the convergence of the K-means algorithm, four K-value selection strategies were investigated. The results demonstrate that different K-values can be selected to lessen the impact for various clustering ranges. Huang et al. established the FPK-mediterranean algorithm to find the most convergent results based on the iterative K-medoids clustering algorithm for immobile points, constructed immobile point equations for each cluster, and solved the set of equations. The findings revealed that this algorithm's clustering efficiency and clustering quality were much higher than those of the conventional K-medoids algorithm [14]. Qi et al. developed the FPK-medoids algorithm [15]. An adaptive kernel fuzzy C-mean algorithm based on the cluster structure was then proposed. The results show that this method has high converging efficiency.

In summary, site archaeological data excavations have used algorithms to process archaeological raw data, and clustering algorithms have been used less in archaeological research. The overview of the research status is shown in Table 1. However, traditional clustering algorithms have significant limitations, and the improvement of clustering algorithms becomes important. Therefore, the study aims to further advance the process of site archaeology by improving traditional clustering algorithms and combining them with archaeological data excavation and classification.

Table 1: Summary of the current state of research

Methods or datasets	Advantage	Disadvantage	Literature
Archaeological Data Sharing Framework, Digital acquisition technology	Semi automated workflow, Open Data, Interoperability	High complexity	Previtali M[7]
Non target data mining	Accelerate overall analysis time	Suitable for mass spectrometry detection of compounds only	Korf Ad[8]
Cubic spline filter, à trous algorithm, RWT	High computational efficiency	Method construction is complex	Tronicke J[9]
Archaeological Information Systems	Integrating multiple technologies	Method construction is complex	Kansa E[10]
Bayesian statistics	Plain language, assuming quantification, clarity, and transparency	The inference efficiency and precision are low	Otárola-Castillo E R[11]
Improved hybrid particle swarm, K-means	Improved convergence speed and accuracy	Method construction is complex	Li Y[12]
K-means, Lris dataset	Improved the selection method for K values	Low clustering accuracy	Yuan C[13]
K-medoids, FPK-mediterranean algorithm	Significant improvement in clustering efficiency and	The method is relatively complex	Huang X[14]

	quality		
An adaptive kernel fuzzy C-mean algorithm	High clustering accuracy	Low clustering efficiency	Qi G[15]

3 Site archaeological data excavation and data generalisation based on attribute induction

3.1 Site archaeological data excavation based on attribute induction

Site excavation data is divided into three areas: heritage management information, heritage information and heritage excavation information. In heritage excavation information, mechanical means such as exploratory boreholes and machine drilling must be used in order to start the excavation, and the complete information can only be retained to the greatest extent possible if all types of data resources are recorded and preserved in detail. These data resources come from the entire process of heritage collection, including documentation, various types of photographs from the excavation, GIS data, exploratory hole stratification data, etc. There is historical and cultural information and physico-chemical information in the heritage information. The physical and chemical information is obtained and recorded with the help of measuring instruments, while 3D models, drawings and photographs provide a more accurate description of all the physical information that is difficult to record in writing, such as shape and colour, and historical and cultural information must be obtained through in-depth identification by archaeologists [16]. Heritage management information contains a number of important activities related to cultural objects and is an important basis for their excavation, archaeology and restoration. Based on the different types of data storage, the study has analysed the attributes of site archaeological data in terms of two types of structured and unstructured recorded data. Unstructured information is divided into two main categories: pictures and videos, both of which are important supplements to textual materials and provide a visual record of the shape, colour and other attribute information of the artefacts. Structured data primarily records textual information on the excavation process and other pertinent information on the management process [17], as well as information on the provenance, characteristics, management, and conservation of cultural items. Table 2 displays a summary of the fundamental details recorded for the items.

Table 2: Partial basic information of cultural relics records

Field	Data type	Length	Description
RegNumber	Integer	50	Total registration number
Onname	Text	No limit	Original name of the collection
Name	Text	No limit	Collection Name
Source	Text	20	Source
Shape	Text	1000	Morphological character
Quality	Text	50	Quality
Grade	Text	20	Collection level
InDate	Text	10	Date of storage
CurrentStatus	Text	No limit	Current situation

The artefacts themselves contain attribute information divided into many aspects of information such as artefact type, texture, period, testing unit, testing person, and the name of the site where they are located. The study summarises the structured information of the artefacts and draws a structured schematic diagram of the artefacts' attributes, as shown in Figure 1.

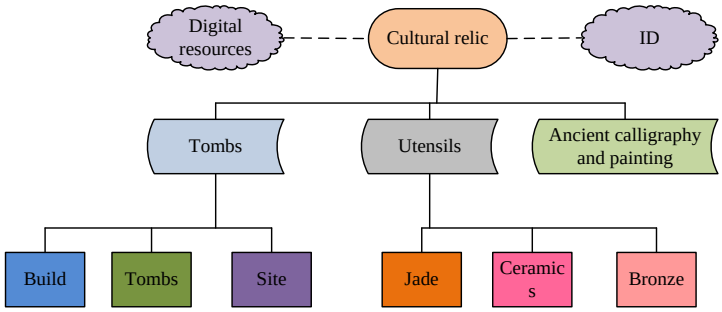


Figure 1: A structured schematic diagram of the attributes of cultural relics

The research is based on identifying the attributes of artefacts from site archaeological data and attribute generalisation through association rules. The association rule method has several uses in data mining and is capable of discovering useful associations in enormous amounts of data. The Apriori algorithm is straightforward and simple to use, making it one of the representative algorithms of association rules. Finding association rules that meet the minimal support and minimal confidence requirements for the frequent item set is the major objective [18]. The joining and trimming operations of the Apriori algorithm correspond to the identification of the frequent item set and the mining of association rules, respectively. In the join operation, the entire dataset is first traversed to obtain the frequent1 itemset L_1, and then the frequent $k-1$ itemset is joined with itself to obtain the candidate itemset C_k, to determine the final frequent k itemset L_k. In the pruning operation, since the candidate set C_k obtained under the concatenation contains elements that do not satisfy the condition, the frequent item set k and L_k must be found by traversing C_k again and deleting item sets smaller than $\min\_\sup$. After completing the above steps, the desired strong association rule is obtained by calculating the confidence of all items to remove items smaller than $\min\_conf$. Figure 2 depicts the Apriori algorithm's flow.

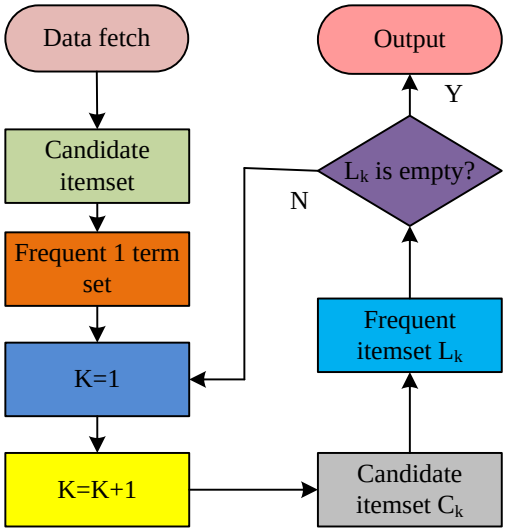


Figure 2: The basic process of Apriori algorithm

The traditional Apriori algorithm tends to obtain a large candidate set and is inefficient. Therefore, the study introduces a boosting degree to improve it. The boosting degree is used as the ratio of the probability of the presence of the posterior term Y to the probability of the occurrence of the posterior term X without X in the presence of the anterior term. Using the lift degree ensures that the association rules mined are positively correlated, as shown in equation (1).

$$Lift(X \Rightarrow Y) = \frac{Conf(X \Rightarrow Y)}{P(Y)} = \frac{P(X \cup Y)}{P(X)P(Y)} \quad (1)$$

In equation (1), the critical value of the lift is 1. $Lift(X \Rightarrow Y)$ is less than 1, the association rule shows negative correlation, when the two item sets exclude each other. There is no correlation with the association rule algorithm when $Lift(X \Rightarrow Y)$ is equal to 1. The association rule is now of scientific interest when $Lift(X \Rightarrow Y)$ is more than 1, which denotes that the association rule exhibits a strong positive connection. It is challenging to successfully discern between the

antecedent and consequent phrases of the rules and to compare various rules when using the lifting degree computation method. Therefore, the study designed a difference-based lifting degree to achieve association rule screening, as shown in equation (2).

$$Lift(X \Rightarrow Y) = \frac{Conf(X \Rightarrow Y) - Sup(Y)}{\max\{Conf(X \Rightarrow Y), Sup(Y)\}} \quad (2)$$

In equation (2), the determination of the influence of on X Y can be made by comparing the probability of the presence of the antecedent Y with the probability of the presence of Y itself when the antecedent X is present. $\max\{Conf(X \Rightarrow Y), Sup(Y)\}$ represents the normalisation factor, which makes the absolute value of $Lift(X \Rightarrow Y)$ less than 1. As $Lift(X \Rightarrow Y)$ is closer to 1, X has great impact on Y , and the closer it is to 0, the more useful the inverse rule of $(X \Rightarrow Y)$ is. The improved Apriori algorithm flowchart is shown in Figure 3.

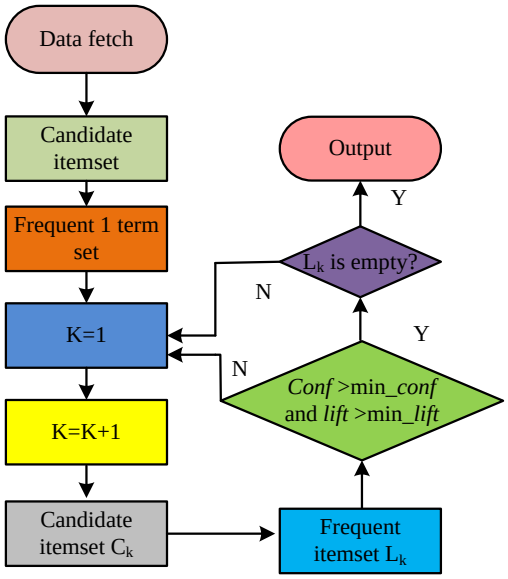


Figure 3: Improved apriori algorithm flowchart

3.2 Data generalisation analysis based on the K-means algorithm

The K-means algorithm is also used in the study to analyse site archaeological data in general. It is based on the introduction of a boosting degree improvement Apriori algorithm. To determine how similar the data are, the classic K-means approach largely analyses the distance between data points and the correlation coefficient between data indicators. In the process of effective clustering and ranking of data with high or low similarity, the correlation coefficient from the previous stage is also used as the basis for judging. In the assignment stage, the data with high similarity to the cluster centres are assigned to the same cluster [19-20]. Figure 4 depicts the fundamental flow of the conventional K-means algorithm.

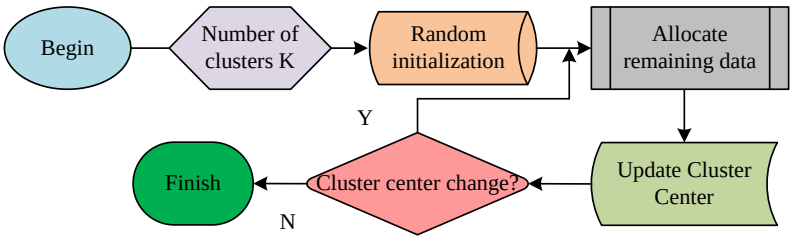


Figure 4: The basic process of traditional K-means algorithm

As can be apparent from equation (3), the Euclidean distance formula is used to determine how complex the data is.

$$d_{x,y} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

In equation (3), x represents the data

($x=\{x_1,x_2,\cdots,x_n\}$) and y is also the data ($y=\{y_1,y_2,\cdots,y_n\}$). As shown in equation (4), the squared value of the error is typically employed in the traditional K-means algorithm as a standard function to evaluate the quality of the data clustering.

$$\begin{cases} C_k=\sum_{x_i\in C_k} x_i/|C_k| \\ E=\sum_{k=1}^{k}\sum_{x_i\in C_k}\|x_i-C_k\|^2 \end{cases} \quad (4)$$

In equation (4), C_k is the centre of clustering, which must ultimately be calculated to minimise the squared error. In this process, the final run of the algorithm ends when both the algorithm convergence and the minimum error squared conditions are satisfied, resulting in the desired classification result. According to equation (4), the traditional K-means algorithm has a relatively high probability of finding data values that fall into a local optimum solution [21-22]. Therefore, a method is devised that allows for the adaptive selection of initial clustering centres based on the distribution of different archaeological data features, in order to avoid local optima as much as possible. Figure 5 depicts the enhanced algorithm procedure.

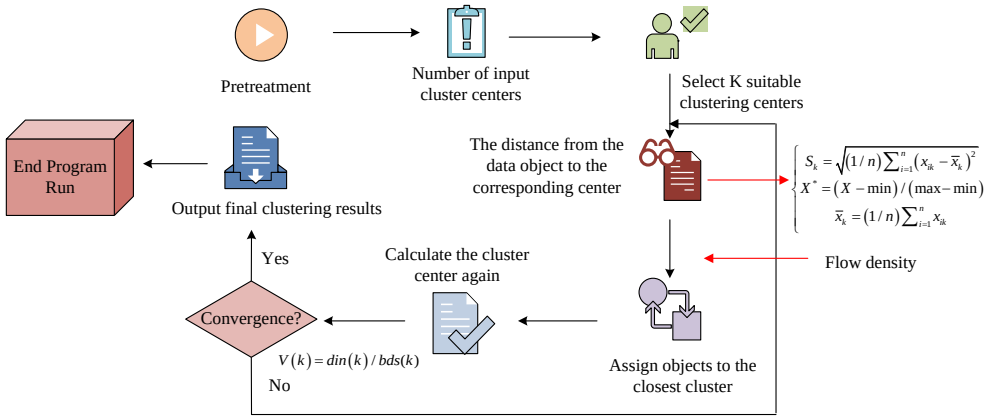


Figure 5: Basic process framework for improving K-means algorithm

The K clustering centres were first chosen by modified K-means algorithm for the input in Figure 4. Then, using the principle of proximity, it divides the data indicators into two groups: those with low similarity and those with high similarity. After a sample has been assigned, a resultant operation must be developed before the convergence of the data is judged and the output is finally obtained. The improved K-means algorithm runs in order to ensure that the gaps between the input data are small, and must be expanded pre-processing, as shown in equation (5).

$$\begin{cases} S_k=\sqrt{(1/n)\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2} \\ X^*=(X-\min)/(\max-\min) \\ \bar{x}_k=(1/n)\sum_{i=1}^{n}x_{ik} \end{cases} \quad (5)$$

X^* represents the greatest data value, the minimum data value, and the normalisation result in equation (5). Following the completion of the pre-processing, the entropy calculation is performed on the data having values between 0 and 1. The archaeological data set is set to $S=\{x_i\,|\,x_i\in R^m, i=1,2,3,\cdots,n\}$, which gives the convergence criterion and metric for clustering between data objects in the algorithm as revealed in equation (6).

$$\begin{cases} C=\sum_{j=1}^{k}\sum_{x_i\in c_j}|d(x_i,o_j)|^2 \\ d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \end{cases} \quad (6)$$

In equation (6), n stands for the overall number of data objects in the set, m for the spatial data's dimensional value, x,y for the attribute vector, and d for the Euclidean distance. $C_j=\{c_{j1},c_{j2},\cdots,c_{jn}\}^T$ represents the clustering centre. c_j represents the classification cluster, the average of all data points represented by o_j in the cluster c_j. The initial number of clusters is denoted by k. For the purpose of to determine the average distance of data objects in the data

set S and to calculate the density of a data object x_i in the set S, the study adds the idea of traffic density into the enhanced K-means algorithm, as indicated in equation (7).

$$\begin{cases} Mdts(S) = \left[2/n(n-1)\right] \cdot \sum_{i,j=1,i\neq j}^{n} d(x_i, x_j) \\ D(x_i) = \sum_{j=1}^{n} u\left(\rho \times Mdts(S) - d(x_i, x_j)\right) \end{cases} \quad (7)$$

In equation (7), ρ represents the radius factor, taking values in the interval [0.5, 2] and the function $u(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$. The mean density is then solved with the mean centres of the data points in the sub-clusters, as shown in equation (8).

$$\begin{cases} MD(S) = (1/n) \times \sum_{i=1}^{n} D(x_i) \\ \overline{x}_i = \left(1/|C_i|\right) \times \sum_{q=1}^{|C_i|} x_q, x_q \in C_i \end{cases} \quad (8)$$

In equation (8), S denotes the data set. $\overline{x}_i$ denotes the mean centres and $|C_i|$ denotes the number of objects clustered within the archaeological data in the cluster C_i. Equation (9) demonstrates that the value of the minimal distance between the mean centres of the data items in each cluster is the distance between clusters over the entire data set.

$$dbs(k) = \min_{1\leq i\leq k} \underset{1\leq j\leq k, i\neq j}{d} (\overline{x}_i, \overline{x}_j) \quad (9)$$

In equation (9), k denotes the quantity of data in the class cluster, and $\overline{x}_i, \overline{x}_j$ is the mean centres of the data points in the clusters C_i and C_j. The intra-cluster distance of the whole set S is the maximum of the intra-cluster distances of k, resulting in the calculation of the intra-cluster distance in the entire set as shown in equation (10).

$$din(k) = \max_{1\leq i\leq k} \left\{ \min_{1\leq j\leq |C_i|} \left\{ \frac{1}{|C_i|-1} \sum_{p=1,p\neq j}^{|C_i|} d(x_j, x_p) \right\} \right\}, x_j, x_p \in C_i \quad (10)$$

In equation (10), $din(k)$ refers to the intra-cluster distance. the gap between data objects in the set that belong to various clusters gets wider the bigger the resulting inter-cluster distance. In order to comply with the above requirements, the study further introduces the clustering effect determination function as shown in equation (11).

$$V(k) = din(k) / bds(k) \quad (11)$$

In equation (11), $V(k)$ is the inter-cluster clustering of the set. The degree of similarity between the data objects in data set S is at its highest when the discriminant function has a minimal value, which implies that the differences between the data items in various clusters also have a maximum value k. In these circumstances, it is possible to choose the cluster value that will have the best clustering effect by choosing an appropriate value as the minimum value of the discriminant function $V(k)$.

4 Effectiveness of site archaeological data excavation and generalisation analysis

The work makes use of association rules to improve the K-means algorithm by generalising qualities to site archaeological data. This uses the Apriori algorithm, which is based on the lifting degree of differences, to filter association rules. First, it is confirmed that the revised Apriori algorithm is valid. This part of the experiment was carried out on different datasets. Three datasets were chosen from the University of California, Irvine (UCI) database for the study: The Agaricus-lepiota (poisonous mushroom) dataset, the groceries dataset, and Voting Records of the United States Congress in 1984. The Agaricus-lepiota dataset contains a total of 8124 samples with 22 different attributes.The dataset describes different sample characteristics of mushrooms, such as color, odor, smoothness, etc.; as well as the sample's categorical labels, edible or poisonous.The Groceries dataset records a daily transaction record of a German supermarket, containing 9835 consumption records, 169 different items, with each purchase transaction corresponding to a list of items purchased by the purchaser.The 1984 U.S. Congressional Voting Records record the results of legislators' votes on 16 different policy issues according to Republican versus Democratic parties, and the dataset contains 435 records.

Table 3 shows that the Institute's revised Apriori algorithm does, however, include lift as a judgement criterion in terms of algorithm running time. The frequent item set obtained by this algorithm only includes the frequent item set of A, which accelerates the algorithm's processing speed. At the same time, the method is more

efficient than the traditional Apriori algorithm in terms of execution as the posterior term only contains the target term in the process of rule formation and no longer requires screening of redundant rules. In terms of the number of association rules, the traditional Apriori algorithm outputs all the rules in the same data set, while the improved Apriori algorithm obtains all the association rules whose back term is A. It has been discovered that the revised Apriori algorithm can eliminate the duplicate rules, resulting in a set of rules that are all tightly connected and positively correlated, increasing the method's effectiveness even more.

Table 3: Comparison of results between apriori algorithm and improved apriori algorithm in three datasets

Dataset name	Support	Confidence level	Evaluating indicator	Number of association rules	Run time/s
Groceries	0.2	0.02	Sup-conf-lift	59	2.83
			sup-conf	341	2.95
Agaricus-lepiota	0.9	0.4	Sup-conf-lift	16	1.76
			sup-conf	718	1.78
Integrated	0.9	0.3	Sup-conf-lift	286	0.36
			sup-conf	2978	0.37

Based on the data in Table 3, further comparison experiments were done, setting the boosting degree to 1 and the minimum support degree to 0.02. Table 4 displays the number of rules found by the two algorithms at various minimum levels of confidence. Table 4 demonstrates an inverse link between the minimal confidence of the Apriori algorithm before and after the enhancement and the number of rules. At a minimum support of 0.1, the traditional Apriori algorithm obtains 3556 rules, but the required target rules are only 583, indicating the existence of 2973 redundant rules. The number of rules for the original Apriori algorithm and the modified Apriori method are 1213 and 270, respectively, at a minimal confidence level of 1, and there are 943 duplicate rules. In this case, applying the traditional Apriori algorithm to attribute summarisation for archaeological data excavation would require a significant amount of time to sift through the redundant rules. In contrast, the study's improved Apriori algorithm is able to cope with the generation of unnecessary rules in order to provide more objective association rules.

Table 4: Number of rules obtained by two algorithms under different minimum confidence levels

Minimum confidence	Improved Apriori algorithm	Apriori algorithm
0.1	583	3556
0.2	583	3556
0.3	583	3556
0.4	583	3556
0.5	547	3378
0.6	369	2203
0.7	314	1958
0.8	297	1732
0.9	288	1520
1.0	270	1213

To further validate the improved Apriori method, the FP-growth algorithm and Node-list Pre-order Size Fuzzy Frequent (NPSFF) Algorithm Based on Fuzzy Association Rule Mining [23] was included to the experiment for comparison. Figure 6 displays the running times of FP-growth algorithm, NPSFF algorithm and before and after the improvement of Apriori algorithm. Figure 6 illustrates how the overall runtime of the four techniques decreased as the confidence level rose. The highest value of the traditional Apriori algorithm is close to 1.0 s, the highest value of the FP-growth algorithm is close to 0.75 s, the highest value of the NPSFF algorithm is close to 0.55 s, and the lowest values of both are 0.22 s, 0.25 s and 0.23 s, respectively, while the highest value of the runtime obtained by the improved Apriori algorithm is only 0.33 s and the lowest value is 0.1 s, which is lower than the other three methods. It can also be found that the improved Apriori algorithm is lower than the other three

methods in the whole process of confidence increase, and the execution efficiency has been significantly improved.

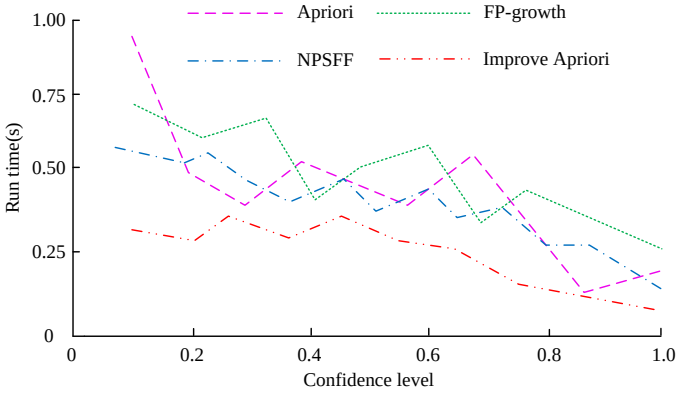
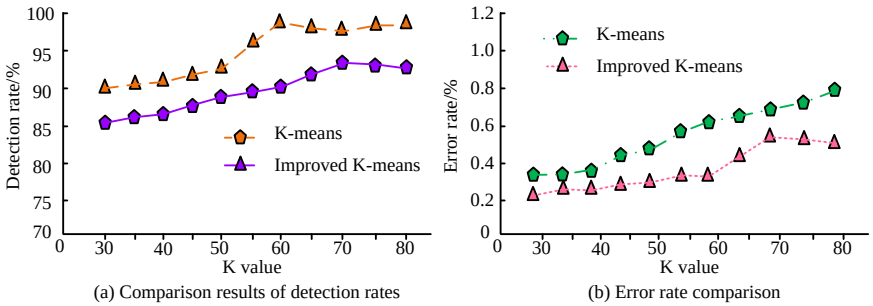


Figure 6: Comparison of running time of FP-growth, apriori algorithm and improved priori algorithm

The study subsequently evaluated the enhanced K-means algorithm after completing the performance validation of the revised Apriori algorithm. Figure 7 displays the findings of the study, which first contrasted it with the conventional K-means algorithm and assessed it using the metrics of data detection rate and error rate. Figure 7(a) and (b) display the outcomes. It has been discovered that as the value of K increases, the detection rate and error rate of the K-means algorithm before and after the improvement show varying degrees of increase. Figure 7(a) shows that the modified K-means algorithm greatly improves upon the classic K-means clustering algorithm in terms of data detection rate. With a detection rate of 98.21% at K=60, the upgraded K-means algorithm's detection rate starts to stabilise. The traditional, enhanced K-means algorithm had a 93.92% detection rate at K=70. Figure 7(b) demonstrates that the modified algorithm has reduced false alarm rate than the conventional algorithm. The enhanced algorithm's error rate at K=70 was 0.402%. The classic algorithm had an error rate of 0.623%. The comparison demonstrates that the optimised research approach can achieve the global ideal answer by having a high detection rate and a low error rate.



(a) Comparison results of detection rates (b) Error rate comparison

Figure 7: Comparison of detection and error rates of K-means algorithm before and after improvement

Figure 8 illustrates how the study contrasts the error of the upgraded K-means algorithm's clustering results with those of the conventional K-means method. The error histogram of the conventional K-means algorithm is shown in Figure 8(a), and the results of the upgraded K-means algorithm's error histogram are shown in Figure 8(b). While there are significant swings in the error range in Figure 8(a), the standard deviation of the K-means method is 4.498. The revised K-means algorithm's standard deviation in Figure 8(b) is 2.537, and a large portion of the error is clustered around 0, with a reduced range of error fluctuation. The upgraded K-means algorithm provides greater classification accuracy and stability as compared to the pre-improvement.

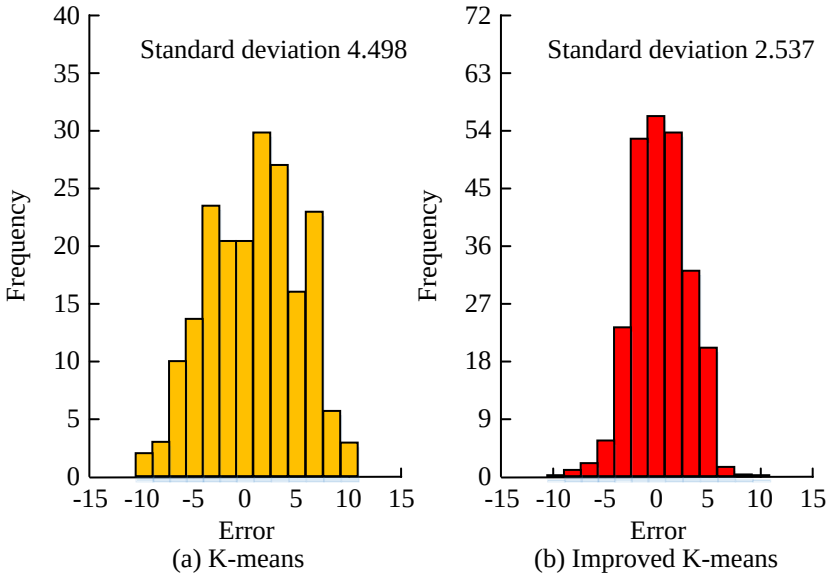


Figure 8: Histograms of prediction error distribution for different algorithms

The study included the Mean Shift clustering method to compare with the enhanced K-means algorithm in order to further demonstrate its superior performance. The dataset used is Flame, which is an artificial dataset and has clusters with ambiguous boundaries, which can verify the effectiveness of the algorithm for noisy data processing. The Flame dataset is a forest fire detection dataset based on aerial images made public by Northern Arizona University and others, and contains 177 different categories. Figure 9 displays the outcomes of the three techniques on the dataset Flame. Figures 9(a) and (b) show that the Mean Shift clustering algorithm and the K-means algorithm both provide outputs that contain a significant amount of noisy and ambiguous data points. And there is a more complex relationship between the real clusters, both fail to accurately identify the basic shape of the clusters, and the clustering effect is poor. In Figure 9(c), the enhanced K-means algorithm accomplishes accurate clustering of this dataset with a better clustering effect and better performance while being minimally impacted by the noise points.

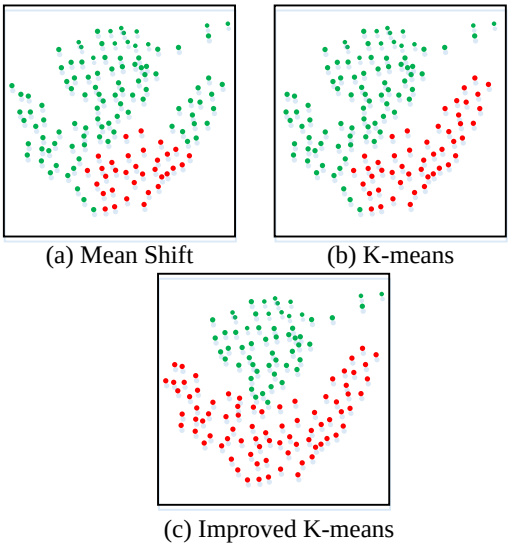


Figure 9: The clustering performance of three algorithms in the artificial dataset flame

The current advanced clustering models Attribute Spectral Clustering (ASC) [24] and Fuzzy decision tree-based clustering algorithm (FDTC) [25] are selected, and Xie Beni (XB) and Davies-Bouldin Index (Davies-Bouldin Index, DB) indicators for evaluation, and the experimental results are shown in Figure 10. As

seen in Figure 10(a), the DB value of the improved K-means algorithm shows a decreasing trend with the increase of the number of iterations and is at the lowest level of 0.07. As seen in Figure 10(b), the XB value of the different clustering methods shows a decreasing trend, and the improved K-means algorithm shows a more obvious advantage of taking the value of the XB, and the clusters of the clustering results have the smallest intraclass distances and the largest interclass distances.

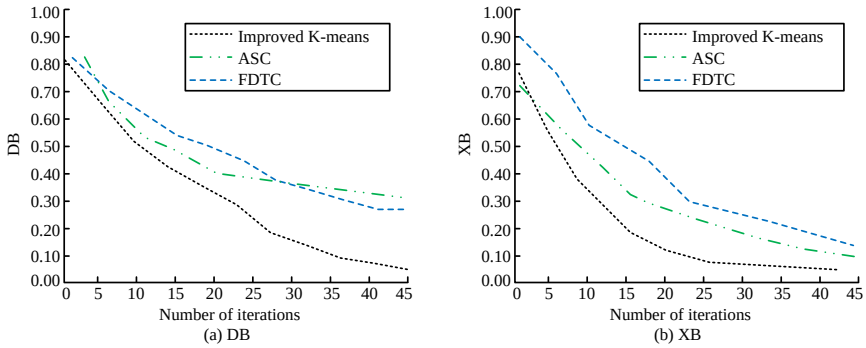


Figure 10: Comparison of statistical results of evaluation indicators of different clustering models

In order to evaluate the scalability and noise tolerance of different clustering methods, the study selects the distributed processing ability, clustering stability and noise point misclassification rate indexes for evaluation, and the experimental results are shown in Figure 11. From Fig. 11 (a), it can be seen that the improved K-means algorithm has the optimal distributed processing ability, and the processing ability evaluation value rises the fastest with the increase of iteration number. As seen in Figure 11(b) and (c), the clustering stability and noise point misclassification rate of the improved K-means algorithm perform optimally. The highest value of clustering stability reaches 92.56%, and the clustering results still have high consistency in the presence of noise. The noise point misclassification rate of the improved K-means algorithm converges to the lowest value of 19.85%, and the lower noise point misclassification rate improves the accuracy of the algorithm.

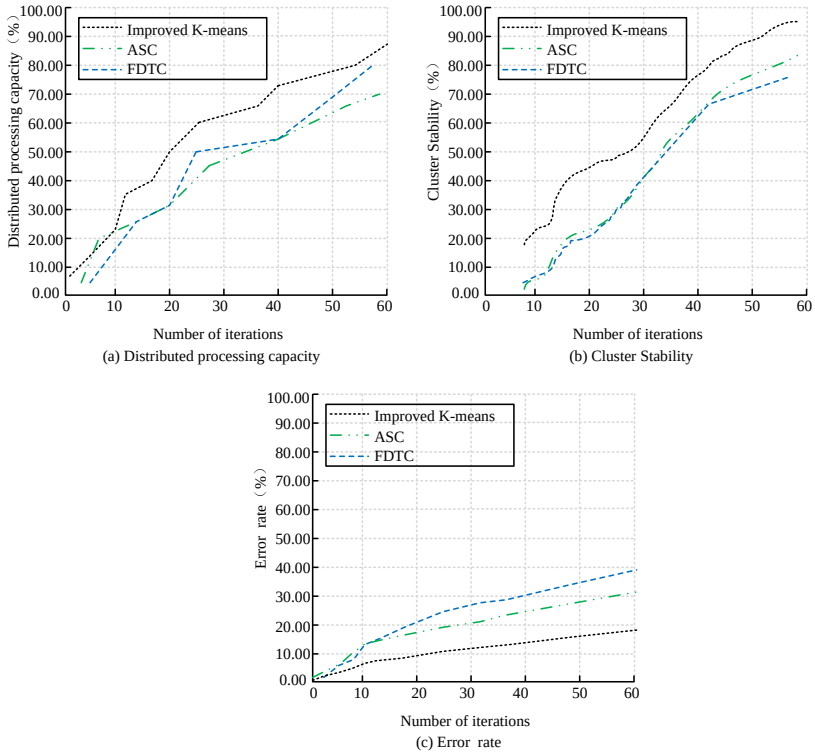
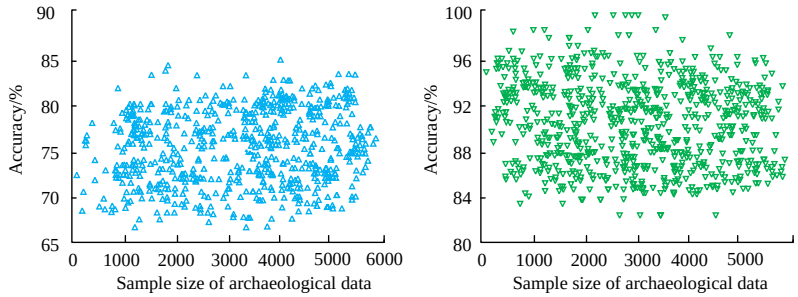


Figure 11: Scalability and noise tolerance of different clustering methods

The study assessed the enhanced Apriori algorithm and the enhanced K-means algorithm and showed that both performed better than expected. They are now jointly applied in practice to validate the results of site archaeological data generalisation analysis. The study selected the digitized resources of Liangzhu Culture Site, one of the southeast Chinese Neolithic culture, Hongshan Site, a noble burial site of the Yue State in Wuxi City, Jiangsu Province, and Sanxingdui Site, a site of the Late Neolithic to Bronze Age in Guanghan City, Sichuan Province, as experimental objects. The heritage data processing of this site was done using the combined technique, and the average results of the data processing for the three different sites are shown in Figure 12. The results of the overall classification accuracy of archaeological data is represented by Figures 12(a) and (b) before and after the combined application of the two methods, respectively. As can be seen from Figure 12, before the joint application of the methods, the classification accuracy of the archaeological data was overwhelmingly below 85% and concentrated at around 75%. After the joint application, the classification accuracy was mostly over 84% and concentrated around 92%, which is a significant increase compared to the pre-application period, proving the effectiveness of the method.



(a) Before joint application

Figure 12: Accuracy of archaeological data classification before and after joint application

5 Discussion

The integration of archaeological digital resources can help researchers access, analyze and compare archaeological resources more conveniently, and understand and excavate the development patterns of ancient civilizations, cultural changes and the laws of human activities. The traditional manual management, electronic documents, shared files and other methods have gradually failed to meet the needs of archaeological work, and the integration of archaeological digital resources needs to adapt to the development of the times and gradually change in the direction of informationization and intelligence. The process of data mining involves a series of steps from data collection to visualization, aiming to accurately search for effective information from massive information. Data mining technology can be used for the description and prediction of target data sets, classification and clustering of data, and identification of abnormal data. Therefore, in the face of massive archaeological data information, the use of data mining can be used to realize the classification and clustering of data, according to the characteristics of cultural relics and sites, mining the connection between different cultural relics and sites, and discovering the trend and cyclical changes of cultural evolution.

In this regard, the study takes data mining as the technical core, adopts the difference enhancement degree improvement Apriori algorithm, and utilizes this method to complete the mining of archaeological data. Meanwhile, the K-means algorithm with adaptive selection of initial clustering centers is designed to further realize the generalization analysis of archaeological data. The improved Apriori algorithm and the improved K-means algorithm perform better in terms of computational efficiency, clustering accuracy, clustering effect, as well as scalability and generalization. The clustering accuracy of the research-designed method is improved by 7.3 percentage points compared to the adaptive kernel fuzzy C-mean clustering algorithm studied by Qi et al [15]. Huang et al [14] designed K-Medoids clustering algorithm based on the immobile point iteration fluctuates in the range of values of the normalized mutual information metrics between 0.7 and 0.8, and the similarity level between clustered labeled clusters is average. In contrast, the research-designed method has reached the optimal value level on DB and XB metrics, and the interclass and intracluster distance status of the clustered clusters has reached the optimal status, and the advantages of the research's improved clustering method are obvious. In conclusion, the study synthesizes the advantages of existing research shown in Table 1 and considers the necessity of information technology in processing large-scale archaeological data. At the same time, the study embodies the research idea of integrating multiple techniques with reference to the current research status quo, and introduces the ideas of difference enhancement degree, adaptive selection, and flow density to improve the traditional Apriori algorithm and K-means algorithm. Compared with the K-means algorithm that introduces intelligent optimization

algorithm, the model complexity is reduced to a certain extent, the model computation is reduced, and the model complexity and computational efficiency are improved.

Comprehensively, the improved Apriori algorithm and improved K-means algorithm proposed by the study perform better in archaeological data mining and greatly improve the management of archaeological data. Based on the analysis and clustering results of data mining, archaeologists can discover more hidden patterns and correlations of cultural relics and sites, providing deeper insights into the study of historical civilizations and promoting the development of the field of archaeology. Considering the development needs of the archaeological industry, on the basis of the research results, future research can introduce more informatization and automation technologies to realize the intelligence of archaeological excavation work, such as the use of computer vision technology to realize the identification and detection of cultural relics, the construction of knowledge maps of cultural relics and sites, and the prediction and extrapolation of cultural relics' historical periods. The development of the field of archaeology will be promoted through in-depth cross-disciplinary cooperation.

6 Conclusion

The large-scale application of digital archaeological data has facilitated the further development of archaeological research. The study summarises the corresponding types of artefact attribute induction for the problem of extensive and complex archaeological data at sites, and applies the improved Apriori algorithm with the K-means algorithm to archaeological data management. This result show that with a minimum support of 0.1, the traditional Apriori algorithm obtained 3556 rules, but the required target rules were only 583, proving that the improved Apriori algorithm can cope with the generation of unnecessary rules. As the confidence level increases, the traditional Apriori algorithm approaches a maximum of 0.1 s, the FP-growth algorithm approaches a maximum of 0.75 s, and the improved Apriori algorithm achieves a maximum runtime of only 0.33 s. In the performance test of the improved K-means algorithm, the data detection rate of the algorithm starts to stabilise when K=60, at which point the detection rate is 98.21%. When K=70, the error rate of the improved K-means algorithm was 0.402%. The error rate for the conventional K-means algorithm was 0.623%. In the error comparison, the standard deviation of the K-means algorithm was 4.498 and that of the improved K-means algorithm was 2.537. With the joint application of the two improved methods, the classification accuracy of archaeological data was mostly over 84%, concentrated around 92%, which greatly improved the management of archaeological data. However, when the proposed method was validated, the study only analysed one selected excavated Neolithic culture site, which may have some errors, and therefore needs to be further optimised by extending the actual validation scope.

Reference

[1] C. Ward, “Excavating the archive/archiving the excavation: Archival processes and contexts in archaeology,” Advances in Archaeological Practice, vol. 10, no. 2, pp. 160-176, 2022. https://doi.org/10.1017/aap.2022.1

[2] I. Sipiran, A. Mendoza, A. Apaza, and C. Lopez, “Data-driven restoration of digital archaeological pottery with point cloud analysis,” International Journal of Computer Vision, vol. 130, no. 9, pp. 2149-2165, 2022. https://doi.org/10.1007/s11263-022-01637-1

[3] S. G. Ortman, and J. H. Altschul, “What north american archaeology needs to take advantage of the digital data revolution,” Advances in Archaeological Practice, vol. 11, no. 1, pp. 90-103, 2023. https://doi.org/10.1017/aap.2022.42.

[4] G. D. Malaperdas, “Practical methods of GIS for archaeologists: Viewshed analysis-the kingdom of pylos example. Geoplanning: Journal of Geomatics and Planning, vol. 8, no. 1, pp. 1-22, 2021. https://doi.org/10.14710/geoplanning.8.1.1-22

[5] D. A. Contreras, Z. Batist, C. Zogheib, and T. Carter, “Matching pragmatic lithic analysis and proper data architecture: The QuARI R shiny database interface,” Advances in Archaeological Practice, vol. 9, no. 4, pp. 299-311, 2021. https://doi.org/10.1017/aap.2021.11

[6] M. A. Jubair, S. A. Mostafa, A. Mustapha, Z. Baharum, M. A. Salamat, and A. Erianda, “A multi-agent K-Means algorithm for improved parallel data clustering,” JOIV: International Journal on Informatics Visualization, vol. 6, no. 1-2, pp. 145-150, 2022. https://doi.org/10.30630/joiv.6.1-2.934

[7] M. Previtali, and R. Valente, “Archaeological documentation and data sharing: digital surveying and open data approach applied to archaeological fieldworks,” Virtual Archaeology Review, vol. 10, no. 20, pp. 17-27, 2019. https://doi.org/10.4995/var.2019.10377

[8] A. Korf, S. Hammann, R. Schmid, M. Froning, H.

Hayen, and L. J. Cramp, "Digging deeper-A new data mining workflow for improved processing and interpretation of high resolution GC-Q-TOF MS data in archaeological research," Scientific Reports, vol. 10, no. 1, pp. 1-9, 2020. https://doi.org/10.1038/s41598-019-57154-8

[9] J. Tronicke, N. Allroggen, F. Biermann, F. Fanselow, J. Guillemoteau, C. Krauskopf, and E. Lück, "Rapid multiscale analysis of near-surface geophysical anomaly maps: Application to an archaeogeophysical data set," Geophysics, vol. 85, no. 4, pp. 109-118, 2020. https://doi.org/10.1190/geo2019-0564.1

[10] E. Kansa, and S. W. Kansa, "Digital data and data literacy in archaeology now and in the new decade," Advances in Archaeological Practice, vol. 9, no. 1, pp. 81-85, 2021. https://doi.org/10.1017/aap.2020.55

[11] E. R. Otárola-Castillo, M. G. Torquato, J. Wolfhagen, M. E. Hill, and C. E. Buck, "Beyond chronology, using bayesian inference to evaluate hypotheses in archaeology," Advances in Archaeological Practice, 2022, 10(4): 397-413. https://doi.org/10.17605/OSF.IO/54F62

[12] Y. Li, J. Qi, X. Chu, and W. Mu, "Customer segmentation using k-means clustering and the hybrid particle swarm optimization algorithm," The Computer Journal, vol. 66, no. 4, pp. 941-962, 2023. https://doi.org/10.1093/comjnl/bxab206

[13] C. Yuan, and H. Yang, "Research on K-value selection method of K-means clustering algorithm," J - Multidisciplinary Scientific Journal, vol. 2, no. 2, pp. 226-235, 2019. https://doi.org/10.3390/j2020016

[14] X. Huang, M. Ren, and Z. Hu, "An improvement of K-Medoids clustering algorithm based on fixed point iteration," International Journal of Data Warehousing and Mining, vol. 16, no. 4, pp. 84-94, 2020. https://doi.org/10.4018/IJDWM.2020100105

[15] G. Qi, W. Guan, Z. He, and A. Huang, "Adaptive kernel fuzzy C-Means clustering algorithm based on cluster structure," Journal of Intelligent and Fuzzy Systems, vol. 37, no. 2, pp. 2453-2471, 2019. https://doi.org/10.3233/JIFS-182750

[16] A. Sabir, H. A. Ali, and M. A. Aljabery, "ChatGPT tweets sentiment analysis using machine learning and data classification," Informatica, vol. 48, no. 7, pp. 103-112, 2024. https://doi.org/10.31449/inf.v48i7.5535

[17] M. H. Santoso, "Application of association rule method using apriori algorithm to find sales patterns case study of indomaret tanjung anom," Brilliance: Research of Artificial Intelligence, vol. 1, no. 2, pp. 54-66, 2021. https://doi.org/10.47709/brilliance.v1i2.1228

[18] F. Moodi, and H. Saadatfar, "An improved K-means algorithm for big data," IET Software, vol. 16, no. 1, pp. 48-59, 2022. https://doi.org/10.1049/sfw2.12032

[19] A. R. Khan, S. Khan, M. Harouni, R. Abbasi, S. Iqbal, and Z. Mehmood, "Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification," Microscopy Research and Technique, vol. 84, no. 7, pp. 1389-1399, 2021. https://doi.org/10.1002/jemt.23694

[20] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," Quality and Quantity, vol. 56, no. 3, pp. 1283-1291, 2022. https://doi.org/10.1007/s11135-021-01176-w

[21] R. Raymond, and M. A. Savarimuthu, "Retrieval of interactive requirements for data intensive applications using random forest classifier," Informatica, vol. 47, no. 9, pp. 35-50, 2023. https://doi.org/10.31449/inf.v47i9.3772

[22] T. Mahmood, and Z. Ali, "Prioritized muirhead mean aggregation operators under the complex single-valued neutrosophic settings and their application in multi-attribute decision-making," Journal of Computational and Cognitive Engineering, vol. 1, no. 2, pp. 56-73, 2022. https://doi.org/10.47852/bonviewJCCE2022010104

[23] T. T. T. Tran, T. N. Nguyen, T. T. Nguyen, G. L. Nguyen, and C. N. Truong, "A fuzzy association rules mining algorithm with fuzzy partitioning optimization for intelligent decision systems," International Journal of Fuzzy Systems, vol. 24, no. 5, pp. 2617-2630, 2022. https://doi.org/10.1007/s40815-022-01308-w

[24] K. Berahmand, M. Mohammadi, A. Faroughi, and R.

P. Mohammadiani, “A novel method of spectral clustering in attributed networks by constructing parameter-free affinity matrix,” Cluster Computing, vol. 25, no. 2, pp. 869-888, 2022. https://doi.org/10.1007/s10586-021-03430-0

[25] L. Jiao, H. Yang, Z. Liu, and Q. Pan, “Interpretable fuzzy clustering using unsupervised fuzzy decision trees,” Information Sciences, vol. 611, no. 8, pp. 540-563, 2022. https://doi.org/10.1016/j.ins.2022.08.077