Fuzzy Similarity K-Type Prototype Algorithm and Marketing Methods

Chuan Tian*, Huan Wan, Ye Wu
Department of Economic Management, Jiangxi Tourism & Commerce Vocational College
Nanchang, 330100, China
E-mail: tianchuan9918@163.com, 13576958158@163.com, 15180560090@163.com
*Corresponding author

Keywords: user feature, marketing approach, improved K-prototypes algorithm, fuzzy similarity matrix, mixed data

Received: November 13, 2024

In the field of user feature segmentation, the currently adopted segmentation methods have the defect of low segmentation accuracy. To address this problem, the study introduces the K-prototypes algorithm for user feature segmentation to improve the segmentation accuracy of user feature segmentation. The study first improves the traditional K-prototypes algorithm using fuzzy similarity matrix. The improved K-prototypes algorithm can effectively select the initial clustering center and fuzzy coefficients and weight coefficients, and pre-set the number of clusters in order to realize the accurate segmentation of user feature. After that, user feature segmentation model is constructed based on the improved K-prototypes algorithm to plan the best marketing methods for users with different characteristics. The study selected 605, 3200, and 684 data objects from the R15, D13, and credit approval datasets as experimental subject. Moreover, it compared the improved K-prototypes algorithm with the fuzzy C-means clustering algorithm and the density peak clustering algorithm in terms of clustering accuracy, root mean square error, mean absolute error, and clustering recall rate to evaluate the performance of the three algorithms. The performance advantages and disadvantages of the three algorithms were evaluated by accuracy, root mean square error, mean absolute error, and recall. The accuracy of the improved K-prototypes algorithm reached 0.9438, which was significantly higher than the other two algorithms. Moreover, the mean square error and mean absolute error of this algorithm were significantly lower than the other two algorithms, indicating that the clustering effect of this algorithm was significantly better than the other two algorithms. The recall of the improved K-prototypes algorithm reached 0.953, and the variation of recall was small, indicating the efficiency of this algorithm in dividing user features. All three algorithms were able to select the correct initial clustering center point for the improved K-prototypes algorithm under different dataset conditions, and the clustering purity of this algorithm was always maintained in the interval of 0.81-0.84. The outcomes reveal that the improved K-prototypes algorithm is able to accurately classify different users according to their characteristic requirements and can plan the best marketing methods for them.

Povzetek: Predstavljen je izboljšan algoritem K-prototipov z uporabo matrike mehke podobnosti za segmentacijo uporabniških lastnosti. Izboljšan algoritem omogoča natančnejšo izbiro začetnih gručnih središč ter določitev parametrov, kar vodi do boljše segmentacije. Model, ki temelji na tem algoritmu, se uporablja za načrtovanje optimalnih marketinških metod.

1 Introduction

With the rapid development of the Internet era, network socialization has gradually become an indispensable part of contemporary people's lives [1]. As a large number of network users carry out various social activities in online social networking, a great deal of data is generated. Among the data generated, most of the personal information of network users is contained [2]. Numerous academics both domestically and internationally have studied this topic in great detail in order to classify these user attributes. S. Y. Wong et al. proposed the use of clustering algorithm for customer segmentation, user segmentation by tracking the stages of the user AIDA model, combining the AIDA model and clustering algorithm to improve the accuracy (AC) of user segmentation [3]. In the field of user feature segmentation, the traditional algorithms in the past could not process and analyze the mixed data, thus the clustering algorithm came into being. One of the most crucial algorithms in data mining is the clustering algorithm. It is able to classify similar objects in a uniform way by the information characteristics of data objects without data samples [4-5]. However, in the actual application of clustering algorithms, the data attributes in the data set are different, both numerical data and categorical data, and even a mixture of numerical data combined with categorical data [6-7]. Therefore, to address the issue of user feature segmentation and the marketing of mixed data, which includes both numerical and categorical data, the k-prototypes algorithm (KPA) was developed. This algorithm combines the advantages of the K-Means and K-Modes algorithms, providing a solution to the aforementioned problems.

At present, KPA still has deficiencies in practical applications. Many scholars at home and abroad have carried out research and improvement to enhance the application effect of KPA. A clustering-based classification technique was proposed by Kuo T et al. The data was clustered using the k-prototypes approach, which used the sine-cosine procedure to determine the weights. According to the findings, the study's suggested algorithm's AC was 92.13% [8]. Y. Shi et al. proposed a marketing strategy based on K-prototypes clustering algorithm (KPCA) in an attempt to achieve clear and comprehensive consumer segmentation. The information entropy weighting was given to the method by particle swarm optimization algorithm as well as hybrid sparrow search algorithm fused with KPCA. The outcomes indicated that the method could effectively perform user feature segmentation [9]. To cluster the recidivism rate of criminals, K. Li et al. proposed a method based on rough sets of probability distributions for the classification. The KPCA was improved by changing the metric of category attributes. According to experimental findings, the enhanced KPCA had an 87.9% prediction AC [10]. R. Bi et al. proposed an additional secret sharing based KPCA in order to alleviate the pressure on resource efficiency. Two servers were utilized to carry out the mixing distance between samples and clustering centers (CCs), and to divide the samples and update the clusters [11]. H. Wang et al. proposed an Intuitive-KPCA by defining the intuitive distribution center of mass and performing a heuristic search for initial prototypes. Combining the two to represent cluster prototypes could help improve the efficiency of mining mixed data [12]. Q. Wei proposed a K-algorithm to simulate student engagement by designing a refined insurance evaluation index system using K-B optimization. Therefore, the proposed strategy of return on investment could ensure the safety of funds and improve the interest of insurance [13].

Aiming at the drawbacks of traditional segmentation algorithms proposed by domestic and foreign researches, the study introduces KPCA. The study uses fuzzy similarity matrix to improve the traditional KPCA and constructs the travel user feature segmentation model. The improved KPCA is utilized to segment user feature and plan the best marketing method. The research aims to improve clustering purity and segmentation AC in mixed data environments through an enhanced K-prototype algorithm. The summary and analysis of related work are shown in Table 1.

2 Methods and materials

2.1 Improved the K-prototypes algorithm based on the fuzzy similarity matrix

In the practical application of clustering algorithm, there are datasets with both numerical and categorical attributes, and it is more difficult to classify different datasets. To solve the problem of harder classification of mixed data and inaccurate user feature segmentation, the study introduces the KPA, which is expected to be utilized to handle large datasets and high-dimensional spaces to enhance the classification performance [14].

Table 1: Summary and analysis of related work

Method	Research field	Advantage	Disadvantage	Author+serial number
Cosine algorithm	Engineering	The accuracy of the algorithm reached 92.13%	Unable to classify data	Kuo T et al. [8]
Particle swarm optimization algorithm and hybrid sparrow search algorithm	Economic	Effectively divide users	Unable to segment consumers	Y. Shi et al. [9]
The method of rough set probability distribution	Criminology	The prediction accuracy is 87.9%	Unable to calculate the recidivism rate of criminals	K. Li et al. [10]
K-prototypes clustering algorithm	Energy management	Real time update of sample clustering	Unable to relieve effective pressure	R. Bi et al. [11]
Intuitive-K-prototypes clustering algorithm	Data science	Unable to partition mixed data	Improve the efficiency of mining mixed data	H. Wang et al. [12]
K algorithm	Education	Unable to calculate student engagement	Ensure financial security and enhance insurance	Q. Wei et al. [13]

benefits

KPA is mainly composed of K-Means and K-Modes algorithms, which mainly deals with the clustering problem of mixed data [15-16]. The algorithm mainly deals with mixed data through the objective function. The calculation formula for the objective function is shown in Equation (1).

$$F(W,Z)= \sum_{j=1}^{k}\sum_{i=1}^{n}\left(w_{ij}\right)^{\alpha}\left(\sum_{r=1}^{p}\left(x_{ir}^{c}-z_{ir}^{c}\right)^{2}+\gamma\sum_{r=p+1}^{m}\delta\left(x_{ir}^{s},z_{ir}^{s}\right)\right) \quad (1)$$

In Equation (1), $F(W,Z)$ denotes the objective function. $\left(w_{ij}\right)$ denotes the elements in the partition matrix. γ denotes the weight scale size. α denotes the fuzzy coefficient. $\delta\left(x_{ir}^{s},z_{ir}^{s}\right)$ denotes the dissimilarity measure. The basic principle of similarity measure is a comprehensive measure to evaluate the degree of similarity between two things. The basic principle of the kernel function is to map the data from the original input space to a high-dimensional feature space. Therefore, it is easier to find linearly separable hyper-planes in this high-dimensional space or to deal with complex non-linear problems. On the basis of the objective function, in an attempt to analyze the affiliation degree of the sample data, mainly by calculating the cluster center of the algorithm. The calculation formula for the cluster center cluster is shown in Equation (2).

$$Z_l=\left\{z_{l1}^{c},z_{l2}^{c},\cdots z_{lp}^{c},z_{l(p+1)}^{s},\cdots,z_{lm}^{s}\right\} \quad (2)$$

In Equation (2), Z_l denotes the cluster center. Among them, the mathematical expression of matrix elements is shown in Equation (3).

$$w_{ij}=\begin{cases}1, d\left(X_i,Z_j\right)\le d\left(X_i,Z_j\right)\\ 0, others\end{cases} \quad (3)$$

In Equation (3), X_i and Z_j denote the sample objects. Based on the sample objects obtained from the calculation, The formula for calculating dissimilarity is shown in Equation (4).

$$d_l\left(X_i,Y_j\right)=\sum_{r=1}^{p}\left(X_{ir}^{c}-Y_{jr}^{c}\right)^{2}, 1\le r\le p \quad (4)$$

In Equation (4), X_{ir}^{c} denotes the subtype attribute of the i th sample. Y_{jr}^{c} denotes the subtype attribute of the j th sample. Calculate the classification attributes of the data sample again, and the mathematical expression of the classification attributes is shown in Equation (5).

$$\delta\left(X_{ir}^{s},Y_{jr}^{s}\right)=\begin{cases}0, X_{ir}^{s}=Y_{jr}^{s}\\ 1, X_{ir}^{s}\ne Y_{jr}^{s}\end{cases} \quad (5)$$

Equation (5) is mainly calculated for sample objects with different categorization attributes. By calculating the resulting objective function and the sample quantity, the specific flow of the KPA for dealing with mixed data is obtained. Its specific flow is shown in Figure 1.

Figure 1 shows the specific flow of KPA. The first step is to input the mixed attribute data and basic parameters, and normalize the data. Then the initial clustering center (ICC) is randomly selected and the sample data are assigned to the clusters based on the dissimilarity measure. Then the fuzzy matrix is calculated based on the assigned data and the CC is updated. Finally, the clustering results are output by checking whether the number of convergences is reached or not, and the process ends after outputting the clustering results. he KPA can handle mixed data with high efficiency and classify different users according to their characteristics. However, it has some problems in the selection of ICC, the pre-setting of the number of clusters, and the selection of fuzzy coefficients and weighting coefficients. Therefore, this study introduces fuzzy similarity matrix to improve the KPA in these three aspects. This is mainly due to the fact that fuzzy similarity matrix possesses the advantage of constructing a coarse particle set and determining the ICC by combining the grain calculation and the maximum-minimum distance method [17-18].

While solving the three aspects of the problem, the algorithm is also able to divide the user's features more accurately. In improving the selection of the ICC of the KPA, it is necessary to avoid the problem of local optimization of this algorithm and to assess the starting point. Therefore, the first step in the improvement process is based on a dense point between the density and the maximum minimum distance as the maximum dense point. The most important step is to enhance the global search capability of this algorithm. Finally, the accurate ICC point is obtained by the maximum and minimum distance. Its specific improvement process is shown in Figure 2.

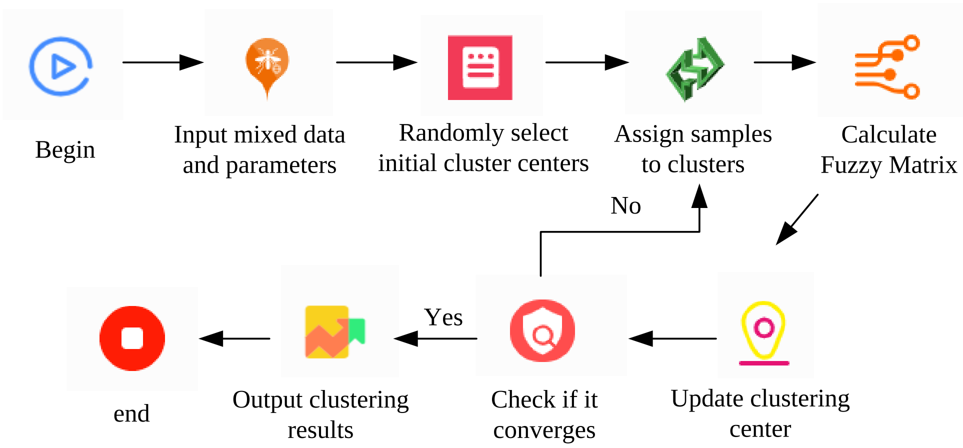


Figure 1: Flowchart of K-prototypes algorithm

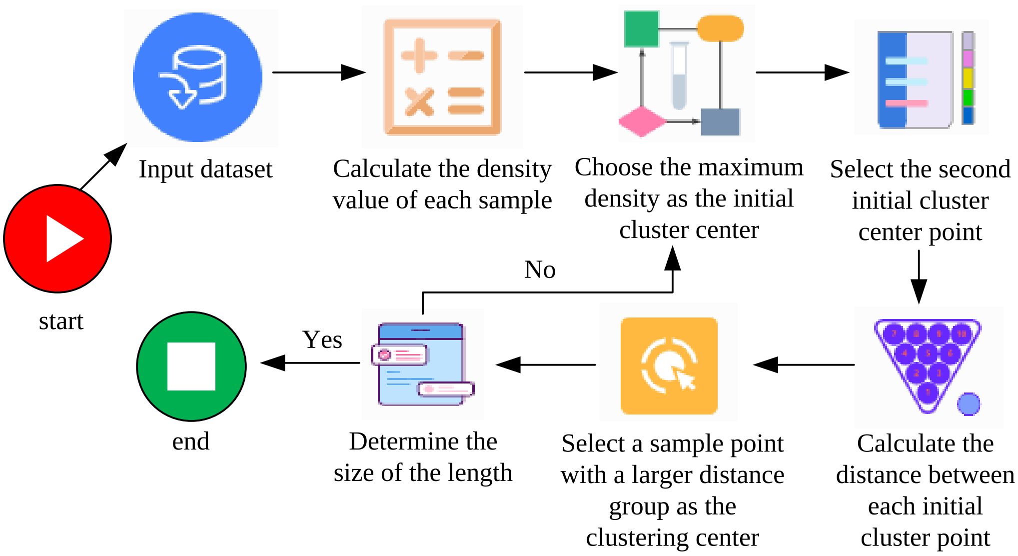


Figure 2: Flowchart of improving the initial clustering center method

The first step in enhancing the ICC aspect in Figure 2 is to compute the total distances between all sample data using the input data. The maximum density value is then chosen as the first ICC point after the average value (AV) of the entire sample data is computed using the sum data. When selecting the second ICC point, it is mainly based on the specific position of the first point. Then the smallest position is selected based on the positions between all the ICC points. The largest position among the smallest positions is selected as the CC point. Finally, determine whether the length of this CC meets the requirement, if it does, then end the process. If it is not satisfied, then continue to select the most suitable CC point. To predetermine the number of clusters, the first step is to calculate the density frequency in the data set. The expression is shown in Equation (6).

$$Dens(u_i) = \sum_{j=1}^{|A|} \left(f_{ij} / n - 1 \right) \tag{6}$$

In Equation (6), f_{ij} denotes the density probability corresponding to the sample object. The density probability of the data is calculated, and the clustering samples are analyzed and integrated based on the data density probability. This process yields the clustering analysis results of the data samples. Then based on the cluster sample set for the sample object, the expression is shown in Equation (7).

$$\begin{cases} Dens(Z) = \underset{X_i \in U}{Max}\left(Dens(u_i)\right) \\ Exemplar_1(u_i) = Dens(u_i) + d(u_i, Z) \end{cases} \tag{7}$$

In Equation (7), Z denotes the largest sample object. $Exemplar_1$ denotes the first cluster sample set. $d(u_i, Z)$ denotes the distance between the sample object and the largest sample object. Then the maximum sample object in the density value is calculated and its expression is shown in Equation (8).

$$d(u_i, Z) = \begin{cases} 0, f(u_i, a_i) = f(Z, a_i) \\ 1, f(u_i, a_i) \neq f(Z, a_i) \end{cases} \tag{8}$$

In Equation (8), $f(u_i, a_i)$ denotes the element value in the sample object. Then the first ICC point is calculated by the candidate set. Its expression is shown in Equation (9).

$$z_1 = \underset{y \in CA_x}{Max}\left(Dens(y) + d(y, Z) - d(y, C_1)\right) \tag{9}$$

In Equation (9), z_1 denotes the first ICC point. C_1 denotes the cluster specimen. Finally, the specimen set is calculated based on the obtained ICC point. The expression is shown in Equation (10).

$$\begin{cases} Exemplar(u) = \\ \underset{u \in U}{Max}\left(Dens(u) + \overset{l}{\underset{i=1}{Mind}}(u, z_i) \right) \\ z_{l+1}(y) = \\ \underset{y \in CA_x}{Max}\left(Dens(y) + \overset{l}{\underset{i=1}{Min}}\left(d(y, z_i) - d(y, u)\right) \right) \end{cases} \tag{10}$$

In Equation (10), z_{l+1} denotes other CCs. In the process of processing the predefined number of clusters, the specific flow of the KPA for this problem is shown in Figure 3.

In Figure 3, the specific process of KPA in pre-setting the number of clusters is as follows. Firstly, the density frequency value corresponding to each sample object is calculated from the input data set. Then the maximum density value is obtained from the calculated density frequency value, and the maximum density frequency value is taken as the sample data. Then based on the sample data the sample dataset with high density and farthest from the sample object is obtained. This data set is used as the first cluster specimen, and the number of CCs is set in advance by seeking the candidate set of ICC through the cluster specimen. Finally, determine whether the number of clusters reserved meets the standard requirements. If it meets the requirements, then end the process, if not, then return to re-select the number of clusters reserved. In the selection of fuzzy coefficients and weight coefficients of KPA, the fuzzy coefficients and weight coefficients are analyzed mainly through the CC points that have been selected. Finally, the fuzzy coefficients and weight coefficients that can be used for user feature segmentation are derived [19-21]. It can handle the fuzzy similarity between data that the K-prototype algorithm cannot handle, and to improve the flexibility and AC of the original K-prototype algorithm. Therefore, the fuzzy similarity matrix can achieve more accurate clustering results faster than the K-prototype algorithm. The novelty of the proposed method lies in changing the mathematical structure of traditional clustering algorithms and optimizing data features through fuzzy matrix clustering. The improved K-prototypes algorithm GK-prototypes algorithm is obtained after improving the above three aspects through fuzzy similarity matrix. The GK-prototypes is able to process mixed data more efficiently and divide them according to the characteristics of different users. Figure 4 illustrates the algorithm's precise structure.

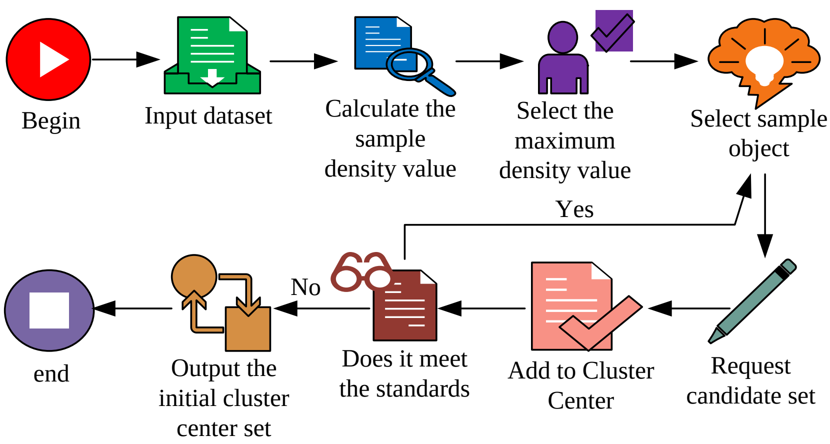


Figure 3: Flowchart of the initial clustering center method for classification

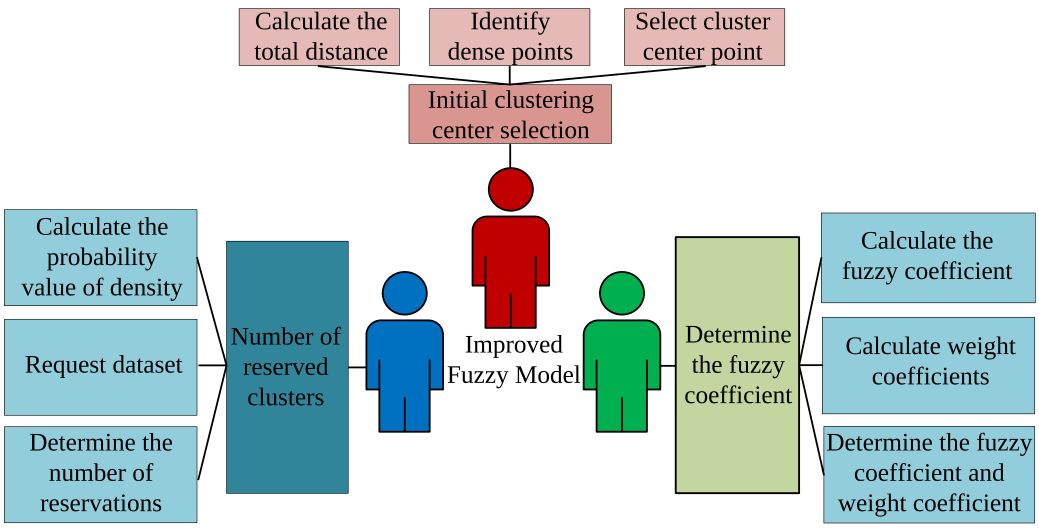


Figure 4: Basic structure diagram of the GK-prototypes algorithm

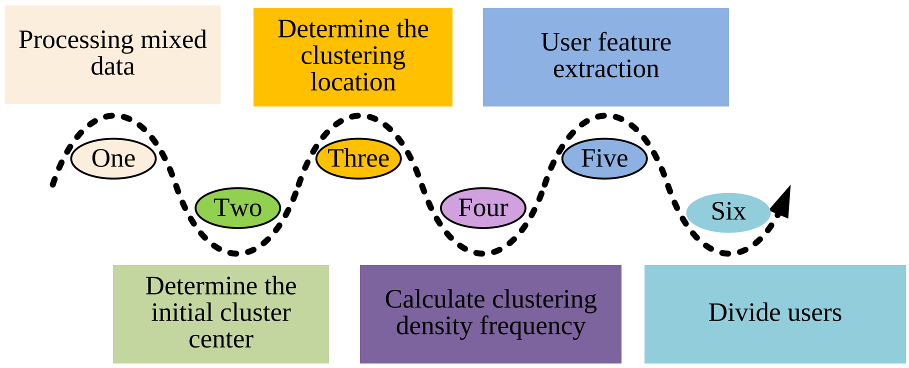


Figure 5: The process of user segmentation using GK phototypes algorithm

In Figure 4, the basic structure of the GK-prototypes algorithm is mainly composed of three parts: the improved ICC, the number of reserved clusters, and the determined fuzzy coefficients and weight coefficients. Each part has a different function, and the GK-prototypes algorithm uses the functions of each of the three parts to process the hybrid data. The main role of the initial class CC is to determine the specific location of each initial clustering point by calculating the data set, which provides a basis for determining the number of reserved clusters at a later stage. Then the density frequency value is calculated based on the location of the initial clustering points. Moreover, the maximum density value is calculated based on the density frequency value, and then the number of clusters is obtained. Finally, on the basis of the number of clusters, the fuzzy coefficient and weighting coefficient are calculated to obtain the finalized fuzzy coefficient and weighting coefficient. To accurately classify the characteristics of users, the basic structure based on GK phototypes algorithm is studied, and users are divided according to their feature requirements. The basic process of classification is shown in Figure 5.

In Figure 5, when using the GK phototypes algorithm for user feature partitioning, the first step is to process the user data and then calculate the ICC points. Second, based on the ICC points, the clustering positions are determined and the clustering density frequency is calculated. Then, the user's features are extracted. Finally, the user's features are divided.

2.2 User feature segmentation and marketing method based on GK-Prototypes algorithm

In the field of user feature planning, the traditional algorithms are not able to accurately divide the features of different users based on different features [22-23]. The GK-prototypes algorithm is able to process mixed data and can accurately divide the feature data. Therefore, for the purpose of improving the AC of identifying the attributes of various users, this study presents the GK-prototypes algorithm. When the GK-prototypes algorithm divides user features, it mainly divides the sample data into several clusters based on similarity. Moreover, the data samples in each cluster have a high degree of similarity. Firstly, the correlation coefficient between user features and optimal features is calculated. The expression is shown in Equation (11).

$$H=\begin{bmatrix} r(X_1,Y_1) & r(X_2,Y_1) & \cdots & r(X_m,Y_1) \\ r(X_1,Y_2) & r(X_2,Y_2) & \cdots & r(X_m,Y_2) \\ \vdots & \vdots & \vdots & \vdots \\ r(X_1,Y_n) & r(X_2,Y_n) & \cdots & r(X_m,Y_n) \end{bmatrix} \quad (11)$$

In Equation (11), $r(X_m,Y_n)$ represents the correlation coefficient between the user features and the feature set. By calculating the resulting correlation coefficient, the distance between the coefficients is calculated. The expression is shown in Equation (12).

$$E(R_i,R_j)=\sqrt{\sum_{k=1}^{m}\left(r(X_k,Y_i)-r(X_k,Y_j)\right)^2} \quad (12)$$

In Equation (12), R_i and R_j denote the coefficient vectors of different users, respectively. Next, using the pertinent coefficient vectors, the distance metric is computed. The calculation process is shown in Equation (13).

$$d_1(X_i,V_j)=\sum_{l=1}^{q}\left|x_{il}-v_{jl}\right|^2 \quad (13)$$

In Equation (13), v_{jl} denotes the AV of the first attribute of the clustered set. In this, the AV of the subtype attribute is calculated based on the AV of the attribute. The expression is shown in Equation (14).

$$\begin{cases} d_2(X_i,V_j)=\sum_{l=q+1}^{m}\delta(x_{il},v_{jl}) \\ \delta(x_{il},v_{jl})=\dfrac{D_{lp}}{D_l} \end{cases} \quad (14)$$

In Equation (14), D_l denotes the sum of distances between all attribute values. D_{lp} denotes the sum of distances between the remaining attribute values. Based on the distance values between attributes and then combined with the distance formula, the distance measurement formula of GK-prototypes algorithm is shown in Equation (15).

$$d(X_i,V_j)=\sum_{i=1}^{q}\left|x_{il}-v_{jl}\right|^2+\theta\sum_{l=q+1}^{m}\delta(x_{il}-v_{jl}) \quad (15)$$

The optimal distance point for user feature segmentation is obtained after the distance measure is calculated by the above steps. To accurately segment different users for their characteristics, the study is based on the GK-prototypes algorithm for characterization segmentation of different users. Its specific process is shown in Figure 6.

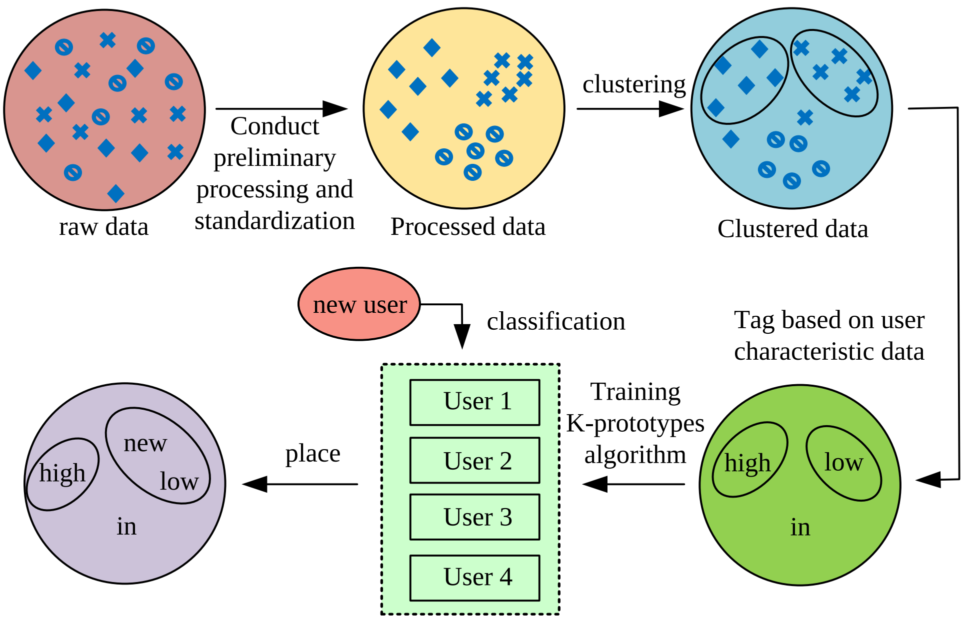


Figure 6: Flowchart of the GK-prototype

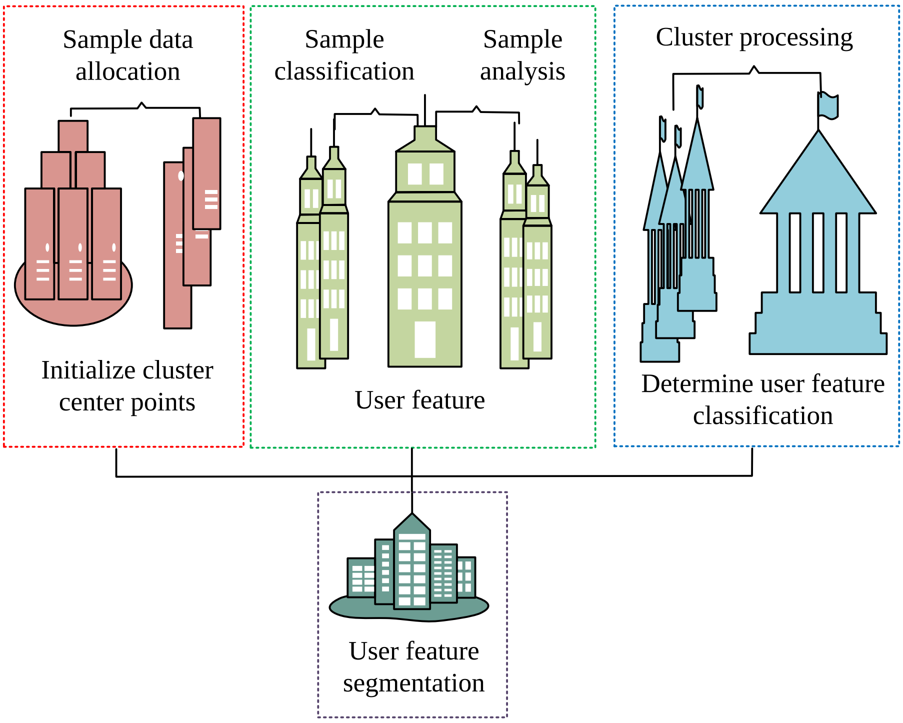


Figure 7: User feature segmentation model created based on the GK-prototypes algorithm

In Figure 6, the GK-prototypes algorithm is first based on the original data collected from the clusters in the division of different user characteristics, and the original data is initially processed and standardized. The processed raw data is divided into different categories based on different features, but its classification is broader. Then the processed data are clustered, and the clustered data have more obvious categories. Then the clustered data is labeled according to the user's features, divided into different feature categories, and the GK-prototypes algorithm is used to classify the user's features. Finally, the features are divided based on different users. To plan the most suitable marketing strategy for the featured users, the study creates a user feature segmentation model based on the GK-prototypes algorithm. Figure 7 displays the model's precise makeup.

In Figure 7, the basic structure of the user feature segmentation model created based on the GK-prototypes algorithm is mainly composed of three parts: ICC, user feature, and determining the user feature classification. The main role of ICC is to process and analyze the sample data. The main role of user feature is to classify and characterize the sample features of the processed data [24-25]. Then the main role of user feature classification is determined to process the user. Finally, the user feature groups are segmented after processing the user sample feature data through three parts. The user feature segmentation model is able to extract and analyze the user feature information. To study and analyze the marketing methods, based on user feature segmentation model, the marketing methods are analyzed. The specific flow of research on marketing methods using this model is shown in Figure 8.

In Figure 8, the specific flow of the model to study the marketing method is as follows. The first step is to calculate the cluster centroids based on the initialized class lake centers. Then the calculated CC value is assigned to the smallest cluster, and the affiliation matrix between the number of samples and clusters is calculated, based on which the affiliation matrix determines the different feature requirements of different users. The cluster centers in the clusters are then recalculated after determining the user feature to determine if there is a change in the number of samples in the cluster. If there is no change, it indicates that the sample is a user determined feature. Finally, it is segmented to determine the marketing strategy. In the event of a modification in the samples included in the cluster, it is necessary to ascertain whether the maximum number of iterations has been reached. If it is reached then it is identified as user feature and it is divided. If it is not able to reach the convergence condition, it is returned and the center of clustering is recalculated so that it reaches the convergence criterion. Moreover, the marketing strategy is determined and the marketing method is investigated.

3 Results

3.1 Performance analysis of the GK-Prototypes algorithm

To verify the effectiveness of GK-prototypes algorithm, the study uses synthetic datasets and datasets from machine learning libraries as experimental datasets. Among them, 20% of the data samples will be used as the test set, and the remaining 80% will be used as the training set. Moreover, to ensure the reliability and repeatability of the experiment, the experimental environment is completed by modifying the content as follows: The Matlab software version 2024a is used. To ensure the AC of the experimental results, the experiment is conducted in the following environment, with a processor equipped with Intel's Core i7-12700/F and a random-access memory size of 32GB. Secondly, the graphics card is equipped with Nvidia's GTX 1660 Super. Finally, the operating system is selected as Windows 11, and the programming language is chosen as Python. Moreover, the traditional algorithms fuzzy C-means clustering algorithm (FCM), density peak clustering algorithm (DPC) are introduced to compare the performance with GK-prototypes algorithm. The performance of different algorithms is mainly evaluated using metrics such as AC, root mean square error (RMSE), mean absolute error (MAE), recall rate (RE), and loss value (LOSS). Its experimental dataset is shown in Table 2.

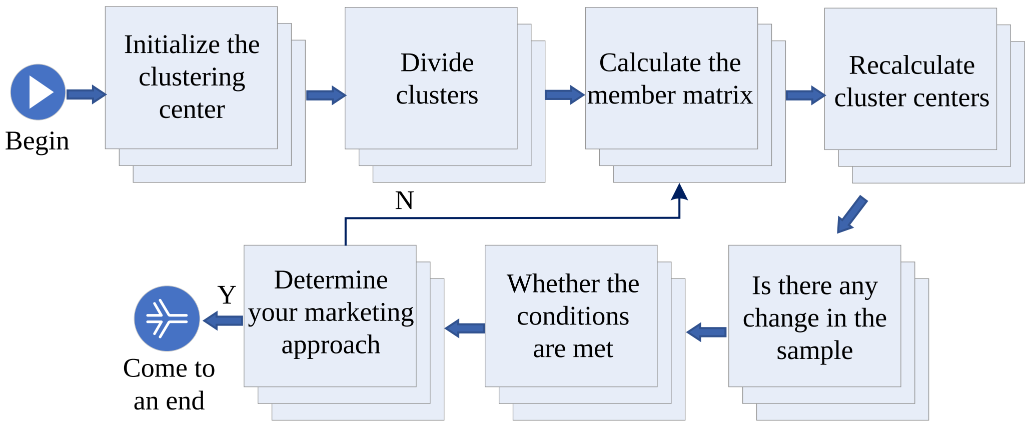


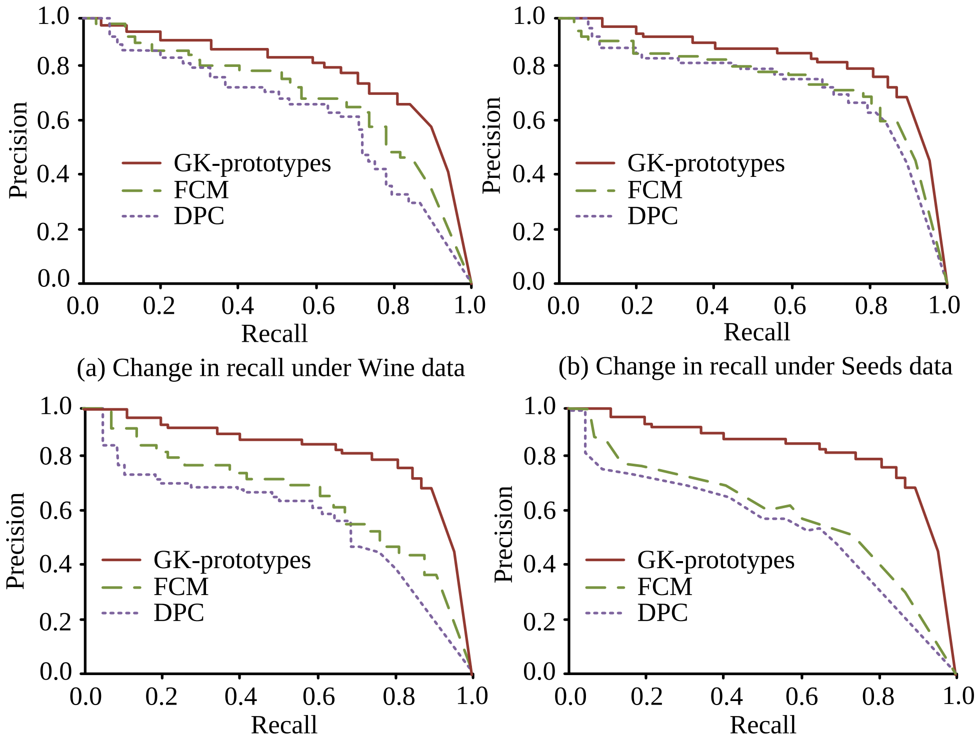
Figure 8: Research on marketing methods based on user characteristic division model

Table 2: Experimental dataset

Data set	Number of objects	The number of numeric attributes	Number of categorical attributes	The true number of clusters
R15	605	2	0	16
D31	3200	2	0	32
Wine	169	11	0	3
Seeds	205	6	0	3
Soybean	46	0	33	4
Statlog heart	268	5	6	2
Credit approval	684	5	8	2

The FCM, DPC, KP and GK-prototypes algorithms are mainly compared in terms of performance in the dataset shown in Table 2. To evaluate the predictive performance of different clustering algorithms under different data conditions, the study compared the REs of three algorithms under the conditions of wine, seeds, soybeans, and R15 dataset, mainly due to the different chemical characteristics such as acidity, ash content, and alcohol concentration measured in wine. These features provide rich information for clustering analysis, allowing algorithms to group wines based on similarity of chemical composition. The seed dataset contains various characteristics such as seed size, shape, color, texture, and biochemical composition. It can also be used for clustering analysis to reveal the similarities and differences between different algorithms. The results of them RE changes are shown in Figure 9.

In Figure 9(a), under the condition of this dataset, the RE of the GK-prototypes algorithm has a relatively flat change amplitude, and the AV of the RE reaches 0.9530. Both the FCM and the DPC show smaller fluctuations in the change curves of them RE under the condition of this dataset. Moreover, their mean values are in the range of 0.8108-0.9541, which is significantly lower than the RE of the GK-prototypes algorithm. This suggests that, in terms of planning for user features under these dataset conditions, the GK-prototypes algorithm's RE is noticeably higher than the other two algorithms. In Figure 9(b), the FCM and the DPC basically converge in them RE change curves under the conditions of this dataset, and neither of them shows large fluctuations. The AV of them RE is maintained in the interval of 0.8904-0.9000. The RE of the GK-prototypes algorithm changes more gently than the other two algorithms. The AV of RE reaches 0.9002, which is slightly higher than the other two algorithms. In Figure 9(c), the RE change curves of the GK-prototypes algorithm and the FCM and DPCs do not show large fluctuations under the conditions of this dataset. Although the change curves of the FCM and the GK-prototypes algorithm basically converge to the same curve, the difference in the AV of their REs is larger. Regardless of the type of dataset conditions, it can be stated that the GK-prototypes algorithm's RE is substantially greater than the other two methods, confirming the algorithm's prediction effect. In Figure 9(d), the GK phototypes algorithm demonstrates a superior RE compared to the other two algorithms on the R15 dataset, with an average RE of 0.9672. The study then analyzes the AC of the three algorithms to assess the ratio of the number of accurate samples to the number of all samples of the three algorithms under various dataset conditions. The results of the variation of their AC are shown in Figure 10.



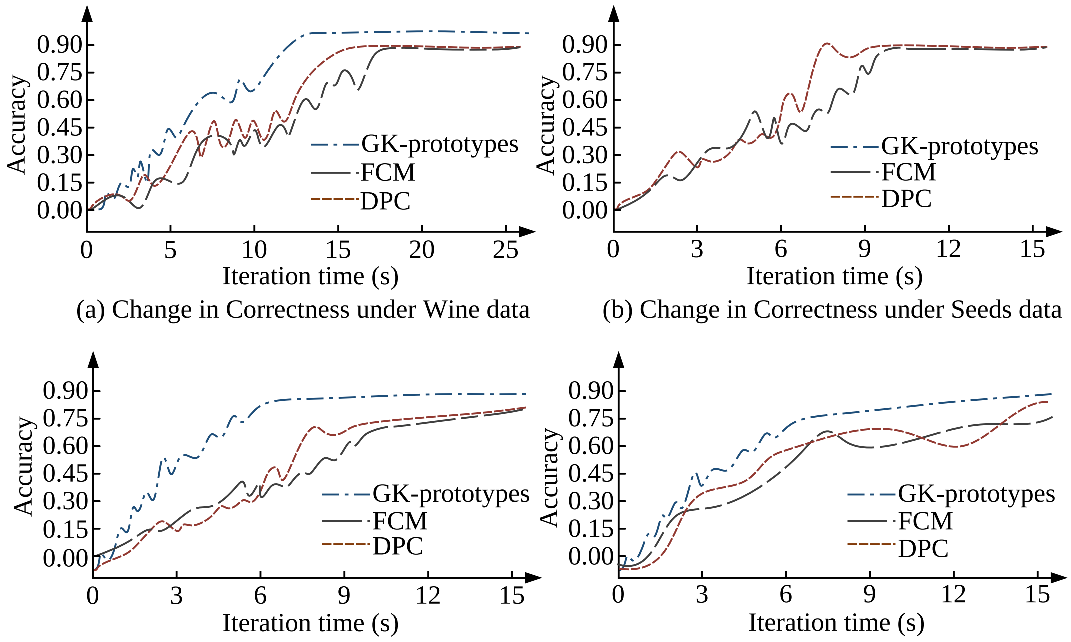
(a) Change in recall under Wine data

(b) Change in recall under Seeds data

(c) Change in recall under Soybean data

(d) Change in recall under R15 data

Figure 9: The results of the recall changes of the three algorithms under different dataset conditions



(a) Change in Correctness under Wine data

(b) Change in Correctness under Seeds data

(c) Change in Correctness under Soybean data

(d) Change in Correctness under D31 data

Figure 10: The results of the accuracy of the three algorithms under different dataset conditions

In Figure 10(a), under the condition of this dataset, when the three algorithms perform user feature segmentation, the AC of the GK-prototypes algorithm for feature sample selection reaches 0.9438. Moreover, the change of its AC does not show a large fluctuation, which indicates that the feature samples segmented by the GK-prototypes algorithm are more consistent with the user's feature requirements. In Figure 10(b), when the three algorithms are dividing the user feature samples, the change curves of the ACs of the FCM and the DPC are consistent, and their ACs are always maintained in the range of 0.8904-0.9000. Whereas, when compared to the other two methods, the GK-prototypes algorithm's AC of 0.9154 is noticeably greater. In Figure 10(c), when the three algorithms perform user feature segmentation under the conditions of this dataset, the AC of GK-prototypes algorithm changes more gently, and the AC reaches 0.9998. The change curve of the AC of FCM basically tends to be the same as that of GK-prototypes algorithm. However, the AC of this algorithm is only 0.7289, which is significantly lower than that of the GK-prototypes algorithm. It can be concluded that, under different dataset conditions, when the three algorithms carry out user feature segmentation, the GK-prototypes algorithm is capable of segmenting feature speech samples that are more in line with the requirements of users. This confirms that this method performs noticeably better than the other two algorithms. In Figure 10(d), under the conditions of the D31 dataset, the GK phototypes algorithm demonstrates a higher selection AC compared to the other two algorithms, with an AC of 0.9761. In addition, to analyze the clustering error of the three algorithms for their categories under the same dataset conditions, the study will compare the RMSE and MAE of the three algorithms. The results of their variations are shown in Figure 11.

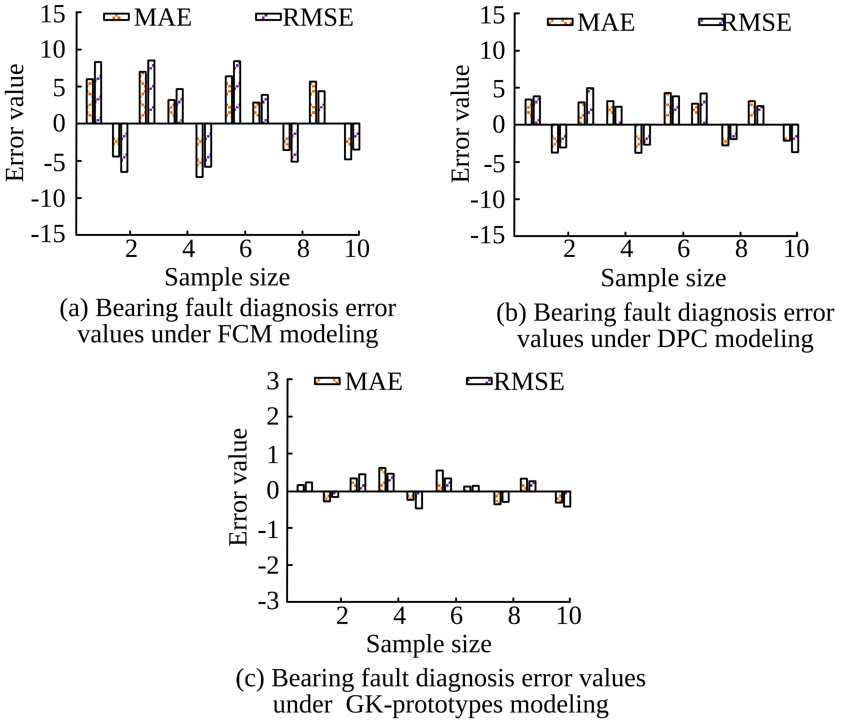
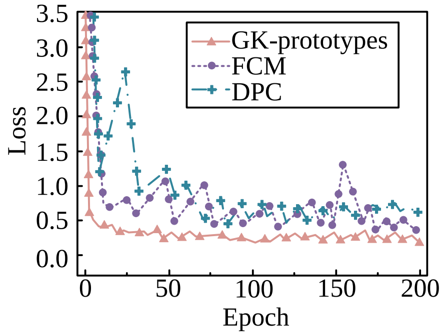
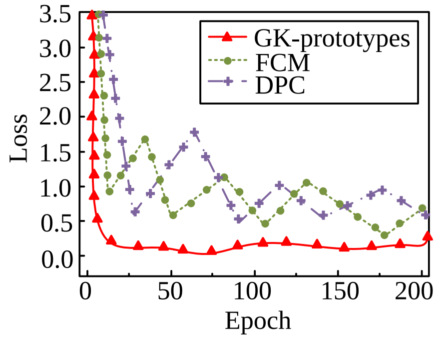


Figure 11: Variation results of MAE and RMSE values of the three algorithms

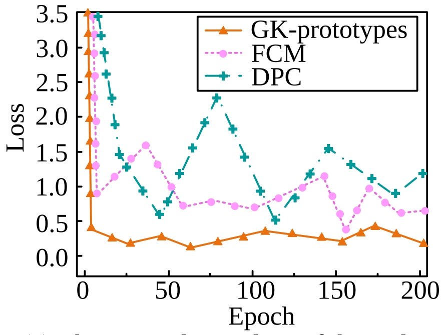
In Figure 11(a), the FCM has a small change in its MAE value under the conditions of this dataset, and the MAE value is always maintained in the interval of 5.554-5.542. The change in the MAE value gradually decreases with the increase in the samples. It displays that the error of the clustering effect of the algorithm under the conditions of this dataset, based on the feature samples of the user feature segmentation, decreases between with the increase in the number of samples. As the samples increases, the RMSE value likewise falls, and the AV of RMSE consistently stays between 7.458 and 7469. In Figure 11(b), under the conditions of this dataset, when the DPC divides the user feature, the change of its MAE value and RMSE value is small, and with the increase of the number of samples. Both MAE and RMSE values do not show large fluctuations, and both remain within the interval of 5.012-5.135. This indicates that the algorithm selects feature samples whose clustering effect is better. In Figure 11(c), when the GK-prototypes algorithm divides the user's feature samples, the changes of its MAE value and RMSE value are relatively flat. Moreover, the AVs of both are always maintained in the interval of 0.512-0.413, and their AVs are significant other two algorithms. This suggests that when it comes to clustering user feature samples, this method performs noticeably better than the other two algorithms. It can be concluded that when using the GK-prototypes algorithm to divide the user's features, it is able to provide the user with feature samples that have the best clustering effect, which verifies the clustering effect of the algorithm. In addition, a comparative analysis was conducted on the loss values of the three algorithms under the conditions of Statlog heart, Credit approval, and R15 datasets to highlight the effectiveness of the GK phototypes algorithm. The ensuing comparison results are displayed in Figure 12.



(a) Changes in loss values of three algorithms under Stat log heart dataset conditions



(b) Changes in loss values of three algorithms under Credit approval dataset conditions



(c) Changes in loss values of three algorithms under R15 dataset conditions

Figure 12: Changes in loss values of three algorithms

In Figure 12(a), under the conditions of the Statlog heart dataset, the average loss value of the GK phototypes algorithm is only 0.379, which is significantly lower than the other two algorithms. In contrast, the average loss values of the DPC algorithm and FCM algorithm are 1.482 and 0.957, respectively. These values are 1.103 and 0.578 higher than the GK phototypes algorithm. In Figure 12(b), the loss values of the GK types, DPC, and FCM algorithms are 0.297, 1.498, and 1.359, respectively. A comparative analysis reveals that the GK types algorithm exhibits a lower loss value in comparison to the other algorithms. In Figure 12(c), under the condition of R15 dataset, the average loss value of GK phototypes algorithm is 0.406, which is significantly lower than the other two algorithms. This finding indicates that the GK phototypes algorithm consistently exhibits lower loss values compared to the other two algorithms under diverse dataset conditions. Then, the performance of GK-prototypes, DPC and FCM are comprehensively compared, and the average performance indexes of the three algorithms are analyzed and compared. The average performance indexes are shown in Table 3.

According to Table 3, the average performance indicators of GK phototypes algorithm are significantly higher than the other two algorithms. Among them, the average MAE value of GK phototypes algorithm is 4.349 and 4.755 lower than DPC and FCM algorithms, respectively. The average AC of this algorithm is 0.9438, which is 0.0438 ($P<0.05$) higher than DPC and 0.0534 ($P<0.05$) higher than FCM algorithm. The outcomes reveal that the optical GK algorithm has significant performance advantages over the other two algorithms.

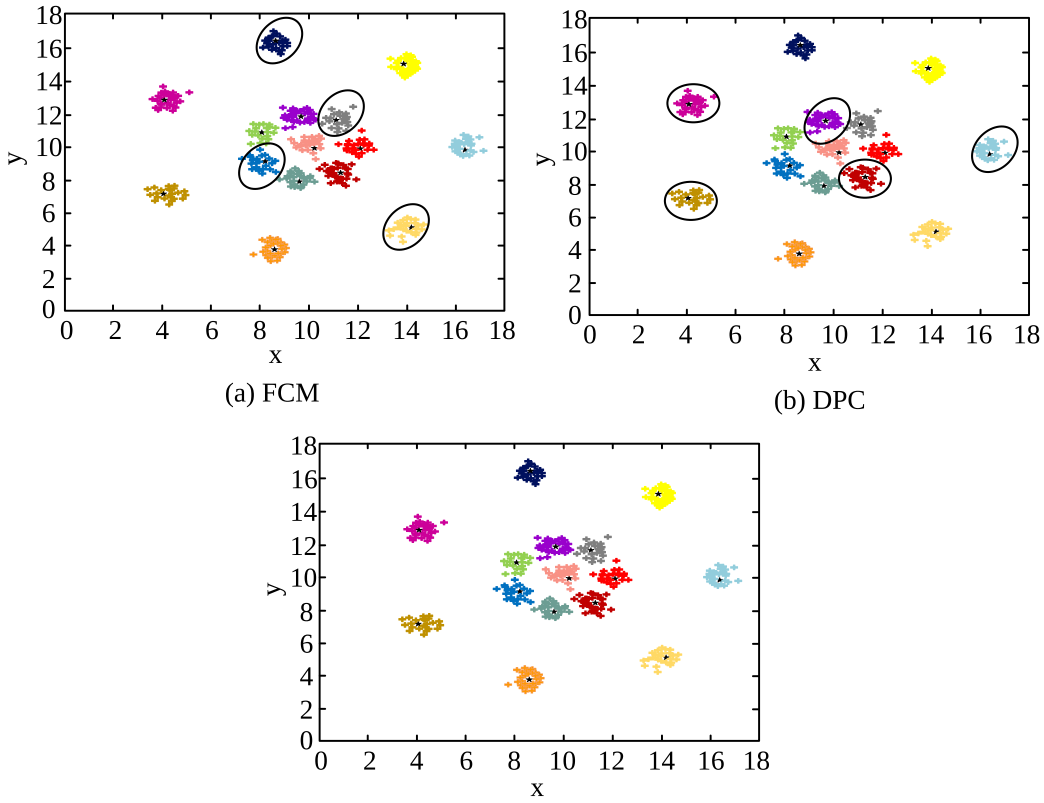
3.2 Empirical analysis of user feature segmentation and marketing methods

605 and 3200 data objects from the R15 and D3 datasets are selected as experimental subjects, mainly due to the fact that the R15 and D3 datasets have 200 data types and are rich in types. The parameters for the number of clusters, the blur coefficient, and the density frequency are set, with the cluster parameter at 4 kb, the blur coefficient at 0–10, and the density frequency at 700 hz. The present study analyzes the clustering effect of disparate algorithms under the conditions of the R15 dataset. The utilization of diverse colors serves to distinguish between distinct class clusters, while black pentagrams are employed to denote centroids. The variation results of its clustering effect are shown in Figure 13.

In Figure 13(a), under the conditions of this dataset, the FCM selects multiple initial centroids but also make four mistakes when selecting the correct initial centroids among different sample class clusters. This indicates that the feature samples selected by this algorithm for user feature segmentation basically meet the user's requirements. However, there are four feature samples that do not meet the use and feature requirements. In Figure 13(b), when the DPC selects the feature samples that best meet the user's requirements among many sample class clusters under the conditions of this dataset, there are four incorrect initial points among the selected initial points. The reason for the appearance of non-initial points is mainly due to the fact that this initial point is in a high-density class cluster, indicating that the correct initial point needs to be selected in a low-density class cluster. In Figure 13(c), the GK-prototypes algorithm does not appear to select the wrong initial point when selecting the most suitable initial point for the user feature criteria among many class clusters. This indicates that the algorithm is able to select the feature samples that best meet the user feature requirements more accurately. It can be concluded that when the three algorithms select the most suitable initial points for user feature requirements among many clusters, the GK-prototypes algorithm is able to accurately select the feature samples that the user prefers and classify them according to the different feature requirements of the user. Then, under the condition of D31 dataset, the clustering effect of feature samples of three algorithms is analyzed when classifying different user features. Figure 14 shows the results of the clustering effect of the three algorithms.

Table 3: Performance indicators of three algorithms

Performance metrics	GK-prototypes	DPC	FCM	p
MAE value	0.4120	4.7610	5.1670	< 0.01
RMSE value	0.5140	5.0340	5.5640	< 0.05
Recall	0.9530	0.8108	0.9541	< 0.01
Accuracy	0.9438	0.9000	0.8904	< 0.01
Loss value	0.3790	1.4820	0.9570	< 0.05



(a) FCM

(b) DPC

(c) GK-prototypes

Figure 13: Clustering effect of three different algorithms on the dataset

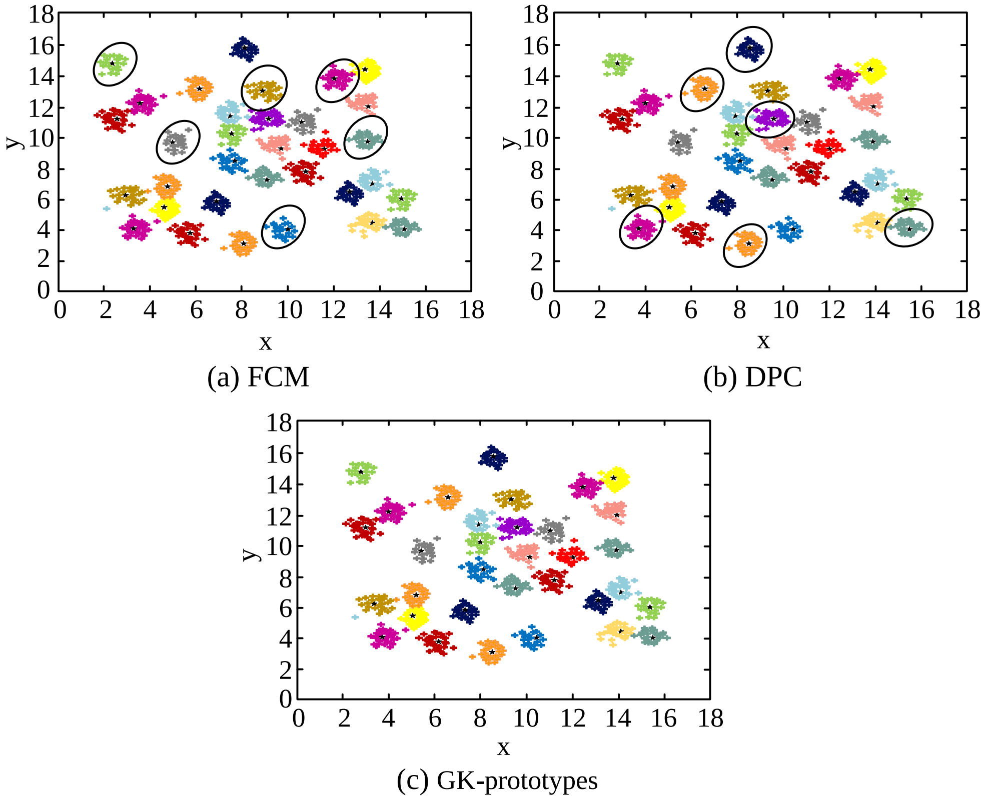
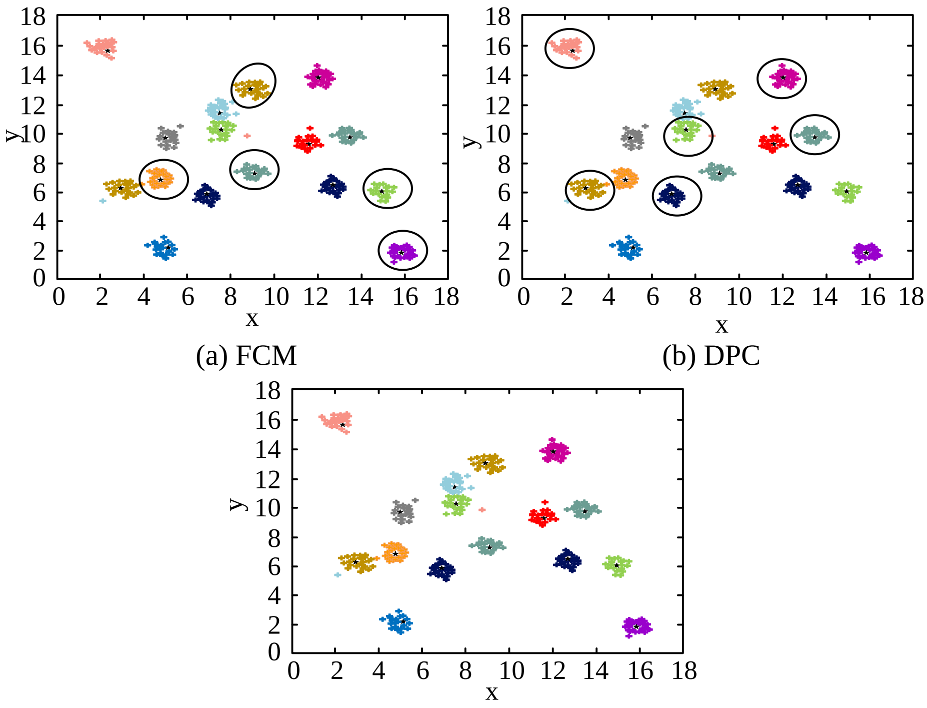


Figure 14: Clustering effect of the three algorithms on the dataset conditions

In Figure 14(a), under the condition of D31 dataset, the FCM is able to select samples that basically meet the requirements of the user feature when selecting initial points in different class clusters, but selects the wrong initial points in individual high-density class clusters. The wrong initial points are the feature samples that do not meet the requirements of the user feature. This indicates that the algorithm is not able to accurately classify the features of different users when the user feature is classified. In Figure 14(b), the DPC, under the condition of this dataset, shows six initial error points when selecting initial points in the unclassified clusters, indicating that the algorithm is able to classify the user feature, although it is able to classify the user's feature. However, it is easy to select samples that do not meet the requirements of user feature when selecting initial points in high-density class clusters. In Figure 14(c), the GK-prototypes algorithm is able to select the initial points that meet the user feature requirements based on the user feature requirements in different density class clusters. It can be concluded that the GK-prototypes algorithm is able to accurately classify the samples according to the user feature requirements when selecting them. Second, under the condition of the credit approval data set, the clustering effect of the three algorithms is analyzed when dividing the characteristics of different users, and the change results of the clustering effect of the three algorithms are shown in Figure 15.



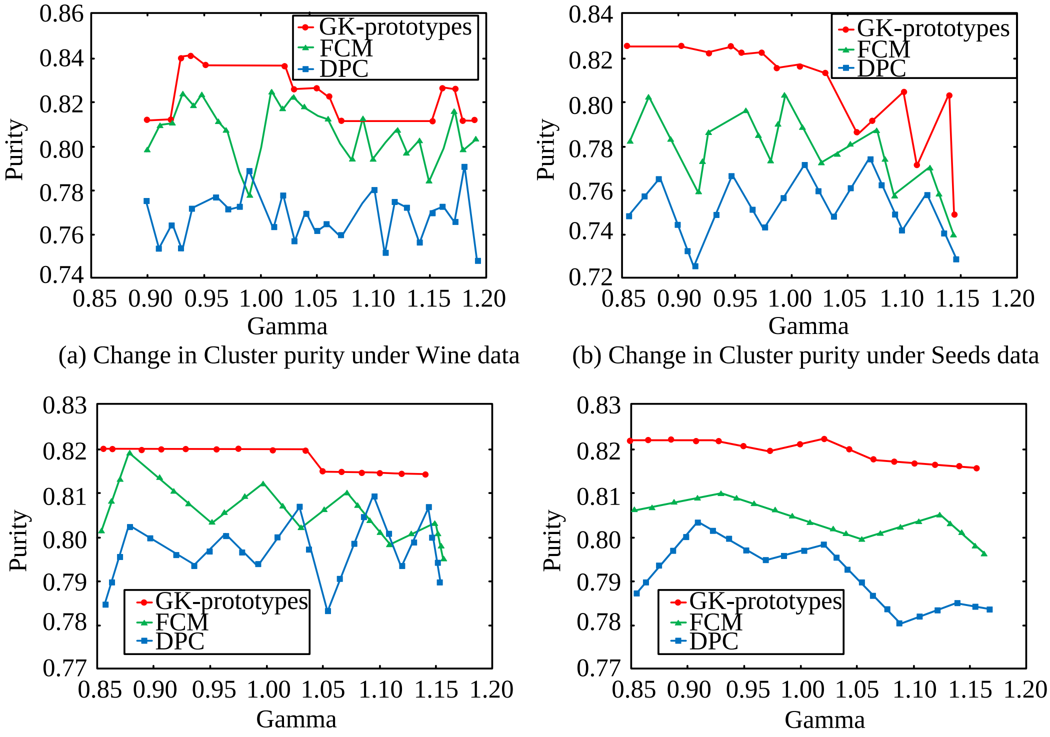
(a) FCM (b) DPC
(c) GK-prototypes

Figure 15: Clustering effect of the three algorithms on the dataset conditions

In Figure 15(a), the FCM algorithm divides the characteristics of users under the credit approval data set, and its identification and clustering effect is poor, and five mis-classifications occur in the identification process. In Figure 15(b), the DPC algorithm is less effective than the other two algorithms, and the number of mis-classifications is significantly higher than other algorithms. In Figure 15(c), compared with other algorithms, the GK-prototypes algorithm can accurately identify and cluster the users' data, and no mis-classification occurs. In addition, to highlight the ability of GK-prototypes algorithm to cluster clusters, the study compares the clustering purity of the three algorithms. Figure 16 displays the outcomes of the clustering purity variation.

In Figure 16(a), the FCM has a larger magnitude of change in clustering purity and a lower AV of clustering purity when the user feature is segmented. While the AV of clustering purity of DPC is higher than that of FCM, but significantly lower than that of GK-prototypes algorithm. The variation of clustering purity of GK-prototypes algorithm is smoother. Moreover, the AV of clustering purity always stays within the interval of 0.81-0.84, and its clustering purity is significantly higher than that of other algorithms. This indicates that under this data condition, when the GK-prototypes algorithm divides the user's features, the clustering purity of its divided samples is significantly higher than that of other algorithms. In Figure 16(b), when the three algorithms divide the user's feature samples under the conditions of this dataset, the clustering purity of the GK-prototypes algorithm reaches 0.83, which is significantly higher than the clustering purity of the other algorithms. Moreover, the change curve of the clustering purity of this algorithm do not show a large fluctuation. In Figure 16(c), under the conditions of this dataset, the clustering purity change curve of the GK-prototypes algorithm basically tends to a straight line, and there is only a small fluctuation. Furthermore, the clustering purity of this algorithm is always maintained in the interval of 0.819-0.820. The clustering purity of FCM and DPC has a large fluctuation in its change curve. This demonstrates that the GK-prototypes algorithm has a far higher clustering purity than the other two algorithms. It also demonstrates that the algorithm is able to accurately classify users' features based on their feature requirements under different dataset conditions. In Figure 16(d), under the condition of the R15 dataset, the clustering purity of the GK phototypes algorithm is notably higher than that of the other two algorithms. The clustering purity of the GK phototypes algorithm reaches 0.82. The objective of this study is to verify the hypothesis that the marketing method proposed by the GK phototypes algorithm has a high user satisfaction rate. To this end, an innovative marketing method is introduced and based on traditional marketing methods. The satisfaction rates of different users are then compared with traditional marketing methods, innovative marketing methods, and the proposed marketing method under the same club conditions. To analyze the application effect of the GK phototypes algorithm, the satisfaction rates of different users with different marketing methods are shown in Table 4.

In Table 4, the marketing method proposed by the GK-prototypes algorithm has a high user satisfaction rate under this club condition. Furthermore, the average user satisfaction rate reaches more than 90%, which is significantly higher than the average satisfaction rate of the other two marketing methods. This indicates that the marketing method proposed by the study can better meet the requirements of different users' characteristics than the traditional and innovative marketing methods.



(a) Change in Cluster purity under Wine data (b) Change in Cluster purity under Seeds data

(c) Change in Cluster purity under Soybean data (d) Change in Cluster purity under R15 data

Figure 16: Comparison of clustering purity results of the three algorithms under different data conditions

Table 4: Different marketing methods for different users

/	/	Customer satisfaction rate		
/	/	VIP member	Regular member	Non member
Traditional marketing methods	Price strategy	87.26%	84.56%	85.69%
	Introduction strategy	89.95%	85.29%	88.46%
	Product strategy	82.39%	87.54%	81.49%
	Promotion strategy	87.63%	83.67%	86.82%
Modern and innovative marketing methods	Price strategy	88.27%	85.59%	86.49%
	Introduction strategy	89.99%	87.25%	88.46%
	Product strategy	84.39%	88.59%	84.46%
	Promotion strategy	88.63%	86.64%	88.84%
Study the proposed marketing method	Price strategy	93.76%	94.59%	91.79%
	Introduction strategy	90.95%	93.89%	91.48%
	Product strategy	92.37%	92.57%	90.79%
	Promotion strategy	90.68%	94.68%	90.89%

4 Discussion

The GK-prototypes algorithm proposed in this study exhibited significant advantages in performance comparison experiments. The GK-prototypes algorithm outperformed the FCM and the DPC in terms of multiple dimensions such as correctness, RMSE, MAE, RE, and other metrics. In particular, the GK-prototypes algorithm was able to maintain a lower error level and a higher AC when dealing with mixed data datasets. Its AC reached 0.9438, which was significantly higher than the other two algorithms. This was mainly attributed to its optimized algorithm design and stronger data processing capability. This result was similar to the conclusion obtained from the study done by O. Pasin in 2023 [26]. In addition, the user segmentation model based on the GK-prototypes algorithm demonstrated strong robustness in coping with the characteristics of different users. Moreover, the clustering purity was always maintained in the interval of 0.81-0.84, which further proved the reliability and usefulness of the algorithm.

In the empirical analysis of user characteristics, division, and marketing methods, the GK-prototype algorithm has the capacity to provide marketing strategies for different groups of people according to their respective hobbies. Furthermore, the algorithm is capable of recommending the most appropriate marketing plans for people with different hobbies. The GK-prototypes algorithm also exhibited significant advantages. The GK-prototypes algorithm outperformed the DPC and the FCM both in the R15 dataset and in the D31 dataset. This was mainly attributed to the fact that the GK-prototypes algorithm was able to segment users according to their feature requirements in an attempt to achieve the effect of having different degrees of clustering in different user features. The GK-prototypes algorithm proposed in the study demonstrated a significant improvement in performance compared to related studies. For example, Hu et al. proposed a modified multi-objective particle optimization algorithm for complex network fuzzy clustering, which improved the clustering convergence speed through the particle swarm optimization algorithm, and the clustering AC reached 91.65% [27]. However, compared with the aforementioned GK-prototypes, there was still a large gap. In addition, the clustering algorithm developed by Hu Scholar could not perform data initialization well, which could also illustrate the limitations of this algorithm. Compared with the algorithms proposed by H. W. Hou et al, the research proposed algorithms demonstrated stronger advantages and utility when dealing with MIXED data sets and diversity data [28]. In addition, G. Zhang et al. proposed a clustering method with local difference privacy. Compared to their study, the GK-prototypes algorithm could help to avoid local optimization and the stability of the algorithm when coding mixed datasets [29]. Furthermore, Y. Li et al. proposed KPCA for glowworm swarm optimization (GSO). Although the study was able to improve the GSO algorithmic method by using the good point set and to design a unified method for the clustering process as well as the categorical data, it lacked practicality in terms of practical application [30]. In contrast, the GK-prototypes algorithm proposes in the study exhibits significant advantages in user feature segmentation and provided the best marketing method for users with different features. However, when splitting datasets with large differences in the distribution ratio of user feature clusters, there is a chance that a small number of user features will be misclassified.

5 Future direction

The enhanced K-prototypes algorithm, as outlined in the present study, is designated as the GK-prototypes algorithm. The GK-prototype algorithm has the capacity to collect the characteristic information of user groups and divide the characteristics of user groups through the clustering analysis of user group behavior data. The formulation of targeted marketing methods and the provision of certain technical support for market segmentation are possible by this process. According to different groups divided by GK-prototypes algorithm, corresponding marketing plans for users. This helps to improve the diversified services of different markets for different user groups, and can improve the satisfaction of different user groups, and improve the marketing efficiency of different user groups of the market.

6 Conclusion

The study introduced the GK-prototypes algorithm, which could solve the problem of the traditional algorithm's low division AC when dividing the user feature. Furthermore, the performance of GK-prototypes algorithm was significantly better than the other two algorithms. When the three algorithms selected the correct initial clustering points in different class clusters, 605 and 3200 data objects from the R15 and D3 datasets were selected as experimental subjects, mainly due to the fact that the R15 and D3 datasets had 200 data types and were rich in types.

When dividing users, the improvement effect is not obvious for datasets with large differences in the distribution ratio of class clusters. Therefore, the clustering effect of large datasets should be improved in the future.

References

[1] J. Liao, A. Jantan, Y. Ruan, and C. Zhou, "Multi-behavior RFM model based on improved SOM neural network algorithm for customer segmentation," IEEE Access, vol. 10, no. 6, pp.

122501-122512, 2022. https://doi.org/10.1109/ACCESS.2022.3223361

[2] Y. J. Han, J. Moon, and J. Woo, "Prediction of churning game users based on social activity and churn graph neural networks," IEEE Access, vol. 12, no. 6, pp. 101971-101984, 2024. https://doi.org/10.1109/ACCESS.2024.3429559

[3] S. Y. Wong, L. Y. Ong, and M. C. Leow, "AIDA-based customer segmentation with user journey analysis for Wi-Fi advertising system," IEEE Access, vol. 12, no. 1, pp. 111468-111480, 2024. https://doi.org/10.1109/ACCESS.2024.3424833

[4] K. Jin, Z. Z. Zhong, and E. Y. Zhao, "Sustainable digital marketing under big data: An AI random forest model approach," IEEE Transactions on Engineering Management, vol. 71, no. 3, pp. 3566-3579, 2024. https://doi.org/10.1109/TEM.2023.3348991

[5] C. H. Hsia, and C. H. Liu, "New hierarchical finger-vein feature extraction method for ivehicles," IEEE Sensors Journal, vol. 22, no. 13, pp. 13612-13621, 2022. https://doi.org/10.1109/JSEN.2022.3177472

[6] E. L. Bonny, and D. Liang, "Exploring sentiment polarity of P2P guests through online user reviews: A multimodal fuzzy approach," IEEE Transactions on Engineering Management, vol. 71, no. 5, pp. 10155-10169, 2024. https://doi.org/10.1109/TEM.2023.3294426

[7] Y. W. Ma, J. L. Chen, L. D. Chen, and Y. M. Huang, "Intelligent clickbait news detection system based on artificial intelligence and feature engineering," IEEE Transactions on Engineering Management, vol. 71, no. 4, pp. 12509-12518, 2024. https://doi.org/10.1109/TEM.2022.3215709

[8] T. Kuo, and K. J. Wang, "A hybrid k-prototypes clustering approach with improved sine-cosine algorithm for mixed-data classification," Computers & Industrial Engineering, vol. 169, no. 4, pp. 164-176, 2022. https://doi.org/10.1016/J.CIE.2022.108164

[9] Y. Shi, S Zhang, S Wang, H. Xie, and J. Feng, "Multiple-perspective consumer segmentation using improved weighted Fuzzy k-prototypes clustering and swarm intelligence algorithm for fresh apricot market," Italian Journal of Food Science, vol. 36, no. 4, pp. 38-57, 2024. https://doi.org/10.15586/IJFS.V2024.2552

[10] K. Li, Z. Wang, X. Yao, J. Liu, H. Fang, and Y. Lei, "Recidivism early warning model based on rough sets and the improved K-prototype clustering algorithm and a back propagation neural network," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 2, pp. 839-851, 2023. https://doi.org/10.1007/Li.2023.12652

[11] R. Bi, D. Guo, Y. Zhang, R. Huang, L. Lin, and J. Xiong, "Outsourced and privacy-preserving collaborative K-prototype clustering for mixed data via additive secret sharing," IEEE Internet Things, vol. 10, no. 18, pp. 15810-15821, 2023. https://doi.org/10.1109/JIOT.2023.3266028

[12] H. Wang, and J. Mi, "Intuitive-K-prototypes: A mixed data clustering algorithm with intuitionistic distribution centroid," Pattern Recognition, vol. 158, no. 4, pp. 62-75, 2024. https://doi.org/10.1016/j.patcog.2024.111062

[13] Q. Wei, "Analysing university student pension insurance using the K-prototypes algorithm and logistic regression model," International Journal of Information and Communication Technology, vol. 24, no. 6, pp. 92-102, 2024. https://doi.org/10.1504/IJICT.2024.138563

[14] G. Duan, S. Yan, and M. Zhang, "A hybrid neural network model for sentiment analysis of financial texts using topic extraction, pre-trained model, and enhanced attention mechanism methods," IEEE Access, vol. 12, no. 2, pp. 98207-98224, 2024. https://doi.org/10.1109/ACCESS.2024.3429150

[15] M. O. Okoye and H. M. Kim, "Optimized user-friendly transaction time management in the blockchain distributed energy market," IEEE Access, vol. 10, no, 4, pp. 34731-34742, 2022. https://doi.org/10.1109/ACCESS.2022.3162214

[16] F. Wu, F. Lyu, J. Ren, P. Yang, K. Qian, and S. Gao, "Characterizing internet card user portraits for efficient churn prediction model design," IEEE Transactions on Mobile Computing, vol. 23, no. 2, pp. 1735-1752, 2024. https://doi.org/10.1109/TMC.2023.3241206

[17] J. Yan, Q. Li, and S. Duan, "A simplified current feature extraction and deployment method for DC series arc fault detection," I IEEE Transactions on Industrial Electronics, vol. 71, no. 1, pp. 625-634, 2024. https://doi.org/10.1109/TIE.2023.3247721

[18] K. M. Hasib, A. Tanzim, J. Shin, K. O. Faruk, J. A. Mahmud, and M. F. Mridha, "BMNet-5: A novel approach of neural network to classify the genre of bengali music based on audio features," IEEE Access, vol. 10, no. 1, pp. 108545-108563, 2022. https://doi.org/10.1109/ACCESS.2022.3213818

[19] M. Li, S. Wang, X. Liu, and S. Liu, "Parameter-free and scalable incomplete multiview clustering with prototype graph," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 1, pp. 300-310, 2024. https://doi.org/10.1109/TNNLS.2022.3173742

[20] Z. Huang, J. Chen, J. Zhang, and H. Shan, "Learning representation for clustering via prototype scattering and positive sampling," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 6, pp. 7509-7524, 2023. https://doi.org/10.1109/TPAMI.2022.3216454

[21] C. Ji, J. Zhou, and F. Kong, "Optimization of prediction intensity of big data clustering algorithm integrated with distributed computing in cloud environment," Informatica, vol. 48, no. 20, pp.

81-94, 2024, https://doi.org/10.31449/inf.v48i20.6054

[22] Z. Zheng, F. Cao, S. Gao, and A. Sharma, "Intelligent analysis and processing technology of big data based on clustering algorithm," Informatica, vol. 46, no. 3, pp. 393-402, 2022. https://doi.org/10.31449/inf.v46i3.4016

[23] W. Chen, and Q. Wang, "Application of improved k-means algorithm in E-commerce data processing," Informatica, vol. 48, no. 11, pp. 147-166, 2024, https://doi.org/10.31449/inf.v48i11.5996

[24] Y. Zhang, and Y. M. Cheung, "Graph-based dissimilarity measurement for cluster analysis of any-type-attributed data," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 9, pp. 6530-6544, 2023. https://doi.org/10.1109/TNNLS.2022.3202700

[25] Y. Shi, S. Zhang, S. Wang, H. Xie, and J. Feng, "Multiple-perspective consumer segmentation using improved weighted Fuzzy k-prototypes clustering and swarm intelligence algorithm for fresh apricot market," Italian Journal of Food Science, vol. 36, no. 4, pp. 38-56, 2024. https://doi.org/10.15586/ijfs.v36i4.2552

[26] O. Pasin, and S. Gonenc, "An investigation into epidemiological situations of COVID-19 with fuzzy K-means and K-prototype clustering methods," Scientific Reports, vol. 13, no. 1, pp. 6255-6274, 2023. https://doi.org/10.1038/s41598-023-33214-y

[27] H. Prasetyo, "Pengelompokan wilayah menurut potensi ekonomi menggunakan modifikasi algoritme fuzzy k-prototypes untuk penentuan target pembangunan desa," Jurnal Teknologi dan Sistem Komputer, vol. 10, no. 1, pp. 46-52, 2022. https://doi.org/10.14710/jtsiskom.2021.14247

[28] L. Hu, Y. Yang, Z. Tang, Y. He, and X. Luo, "FCAN-MOPSO: An improved fuzzy-based graph clustering algorithm for complex networks with multiobjective particle swarm optimization," I IEEE Transactions on Fuzzy Systems, vol. 31, no. 10, pp. 3470-3484, 2023. https://doi.org/10.1109/TFUZZ.2023.3259726

[29] G. Zhang, X. Chen, H. Wang, R. Zhai, and Z. Ma, "K-Prototypes clustering method for local differential privacy," Journal of Computer Applications, vol. 42, no. 12, pp. 3813-3821, 2022. https://doi.org/10.11772/j.issn.1001-9081.2021101724

[30] Y. Li, Z. Ni, and W. Zhou, "Clustering algorithm for mixed attributes data based on glowworm swarm optimisation algorithm and K-prototypes algorithm," International Journal of Bio-Inspired Computation, vol. 18, no. 2, pp. 105-113, 2021. https://doi.org/10.1504/IJBIC.2021.118095