Business Data Analysis Based on Kissmetric in the Context of Big Data

Kan Wang
College of Computer Engineering, Henan Institute of Economics and Trade, Zhengzhou, China

***Abstract*—The kissmetric data analysis model can be used for the analysis and research of business data, and the focused research method in this model is cluster analysis. To realize the effective application of Kissmetric data analysis model, the focused method is improved in the experiment. An improved hierarchical clustering algorithm generated by splitting stage and merging stage is proposed in the experiment, and then the algorithm is combined with density clustering method while considering noise point processing to achieve automatic determination of clustering centers and improvement of clustering effect. In different dimensions, the highest F-measure index and ARI values of the hybrid clustering method are 0.997 and 0.998, respectively. In different numbers of classes of the dataset, the highest F-measure index and ARI values of the hybrid clustering method are 1.000 and 0.999, respectively. The mean accuracy and mean-variance were 95.94% vs. 5.89%, 94.72% vs. 0.57%, 89.72% vs. 4.97%, 87.45% vs. 5.53%, 93.83% vs. 5.76%, and 88.43% vs. 5.40 %, respectively. The mean and mean squared deviation of hybrid clustering method's accuracy was 89.71% vs. 6.17% and 88.85% vs. 0.33% when dealing with the real datasets 7 and 8, respectively. The quality and stability of the clustering results of the hybrid clustering method are better. Compared with other clustering methods, the accuracy and stability of this method are higher and have certain superiority.**

***Keywords*—Big data; kissmetric; data analysis; density clustering; hierarchical clustering**

I. Introduction

In the context of big data, by collecting business data, users can analyze the data according to the actual needs and get the latent customer behavior pattern behind the big data [1]. The business data analysis methods mainly include the list method, statistical method, cluster analysis method, etc. [2]. The list method is to record and process the data results by certain rules by applying tables, which should be designed to meet clear and precise correspondence [3]. The statistical method can collect and organize data with characteristics from microstructure and use suitable statistical methods to organize the data and explore the hidden laws of the macroscopic nature behind the data [4]. Clustering analysis is a simple and efficient way to summarize the data on a website to determine whether there is a correlation between things [5]. Kissmetric is a data analysis model that can be used for customer engagement. The data analysis of this platform can help users understand customer engagement, analyze product performance, and determine whether the customized marketing plan is effective [6-7]. However, Kissmetric contains a large amount of data, and users may not be able to quickly obtain effective information from it. Therefore, it is also necessary to analyze these data in order to improve the application effectiveness of the data analysis model. Research has shown that clustering algorithms have good applications in the analysis of business data [5]. It can be used for data analysis in the Kissmetric data analysis model. However, traditional clustering algorithms require a predetermined k value when applied. The setting of k value is easily influenced by subjective factors, resulting in significant differences in clustering results. In order to reduce the impact of subjective factors, some scholars proposed hierarchical clustering analysis [8]. This method only requires fewer or no parameters, making it highly flexible. Therefore, hierarchical clustering is selected as the main algorithm for business data analysis in this study. However, this method requires a large amount of computation and cannot trace back to the intermediate clustering process. In this experiment, an improved hierarchical clustering algorithm is proposed, which is generated by the split phase and the merge phase. Then, while considering noise point processing, the algorithm is combined with density clustering method to achieve automatic determination of clustering centers and improve clustering performance. It is hoped that the improvement of the method can improve the application effect of clustering methods in the Kissmetric data analysis model. It is hoped that this data analysis model can help users analyze business data and obtain the information hidden behind the data, which can be used for formulating enterprise development goals and directions.

The article is mainly divided into five parts. Firstly, there is an introduction as the background of the article. Then there is a literature review, which discusses the existing methods and serves as the literature basis for selecting methods in the experiment. The third part is the establishment and improvement of methods. The fourth part is the performance analysis of the method. The last part is the conclusion, which is a summary of the entire text.

II. Review of the Literature

The era of big data requires us to process the data efficiently so as to solve the problems brought by the data. When dealing with complex business problems, the laws behind the data need to be mined to uncover the business value represented by different data. The analysis of business data mainly consists of searchable data analysis and model selection analysis. When the data in search web pages are disorganized, searchable data analysis can be achieved by using mapping, generating tables and equation fitting [9-10]. The search data analysis can be used to obtain the potential business value behind the data, based on which a suitable business model can be selected to promote the long-term development of the company. Kissmetric is a data analysis model that can be used for customer engagement [7-8]. The main components of the Kissmetric data analysis model include the visitors to the web pages, the features used to describe the user information, and the events and their attributes. The hierarchical clustering method is the key method used in the Kissmetric data analysis model for statistical analysis of data, which is divided into three steps. The first step is to count and divide the number of customers, the second step is to count and divide the types of products, and the third step is to count and divide the platforms. Hierarchical cluster analysis can help users to analyze business data and get the hidden information behind the data, which can be used for the formulation of business development goals and development directions.

Kissmetric is an automated customer engagement data analysis model based on a hierarchical clustering approach, and clustering algorithms are an important part of data mining. Many classical clustering algorithms have been proposed, such as K-means, DBSCAN, Gaussian mixture models, spectral clustering, non-negative matrix decomposition-based clustering, and graph-based clustering. Many optimizations have been made on the traditional algorithms to enable more efficient and accurate data mining using clustering analysis. For example, Xi W A et al. proposed a memetic algorithm with adaptive inverse K-means operation for data clustering, and the performance of the method was evaluated on a series of data sets and compared with related algorithms, and the experimental results showed that the algorithm generally provides superior performance and outperforms related methods [11]. Optimization improvements to K-means clustering methods can be applied to multidisciplinary research. Liu S et al. proposed a self-guided reference vectors (RVs) strategy for decomposition-based evolutionary algorithms in multi-objective optimization to extract RVs from a population using an improved K-means clustering method [12]. A neighborhood network layout scheme based on the unsupervised K-means clustering algorithm and the contour index method has been proposed to determine the number of effective data aggregation points (DAPs) required for different smart meter densities and to find the optimal deployment locations of DAPs [13]. Qin X et al. proposed a machine learning K-means clustering algorithm to select interpolative separable density fitting (ISDF) interpolation points, and K-means algorithm can significantly reduce the computational cost of selecting interpolation points by nearly two orders of magnitude, thus speeding up the ISDF-based Hartree-Fock exchange computation by a factor of 10 [14]. In the study of enterprise business model change, Dressler M et al. used clustering algorithm to data mine different business models and then used PCA analysis to generate two classes of business models, which provided a better classification method for business model expansion [15]. In the development of SMEs, some scholars use two-step clustering method to analyze the financial data of enterprises to get the optimal number of clusters, and then use fuzzy clustering to analyze the business scale of enterprises, and the optimal clusters are obtained after verification [16]. For the investment selection of enterprises, Gubu L et al. introduced the Markowitz model based on the K-means algorithm for estimating the covariance matrix as well as the mean vector. The method was experimentally demonstrated to have good robustness in processing a large amount of data and can effectively use outliers for data analysis [17]. Clustering algorithm can optimize the parameters of the machine learning model, and the optimized method can be used to cluster the customer preferences and predict the market trends. Clustering-based analysis can uncover the patterns and trends behind customer behavior, which is important for business scaling [18].

Based on the above study, clustering algorithm has good application in the analysis of business data. In order to better improve the application of Kissmetric data model, the focused clustering method in Kissmetric data model is improved in this experiment. An improved hierarchical clustering algorithm generated by splitting stage and merging stage is proposed in the experiment, and then the algorithm is combined with the density clustering method while considering the noise point processing to achieve the automatic determination of clustering centers and the improvement of clustering effect. It is hoped that the improvement of the Kissmetric data model focus method can help users to improve the effectiveness of business data analysis.

III. BUSINESS DATA ANALYSIS BASED ON KISSMETRIC

A. *Business Data Analysis Method based on Improved Hierarchical Clustering*

The focus of Kissmetric data analysis model is the data analysis method based on cluster analysis. In order to improve the application of Kissmetric data analysis model, the focus of this experiment is improved for its focus. K-means algorithm is a representative clustering method in cluster analysis, and is the basis of other clustering algorithms. In the practical application, it is necessary to determine the number of data object classes, divide the points with different object attributes into different cluster classes according to the principle of closest distance, then use the averaging method to calculate the center of mass of each cluster class, and then reassign the center of mass according to the actual calculation results, and finally ensure that the center of mass moves below the set threshold, and Fig. 1 shows its iterative process.

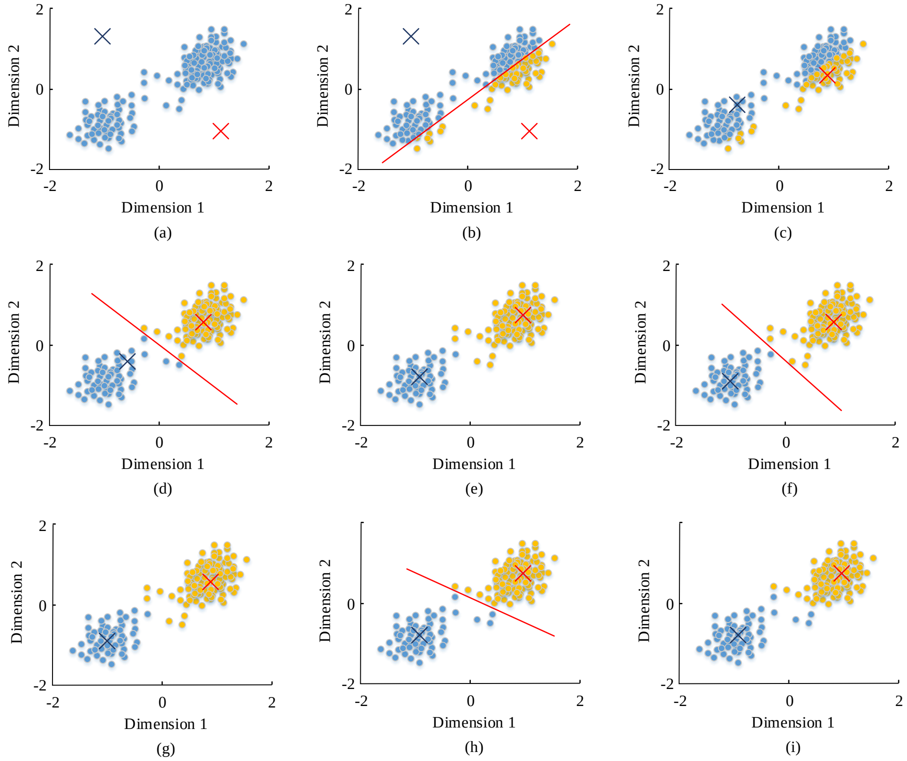


Fig. 1. Iterative process of K-means algorithm.

K-means algorithm requires a pre-specified k-value, but it is susceptible to the influence of subjective factors, which leads to large differences in the results of clustering. To reduce the influence of subjective factors, some scholars propose hierarchical cluster analysis method, which is mainly divided into cohesive and split hierarchical cluster analysis methods. Only fewer parameters or no parameters are required in the hierarchical cluster analysis method, which is more flexible and can be used to divide different types of data according to the different levels of data objects. Therefore, hierarchical cluster analysis method is chosen as the main algorithm for business data analysis in this study, but this method is computationally intensive and cannot retrace the intermediate clustering process, so a new hierarchical clustering method consisting of two phases, merging and splitting, is proposed in this experiment. The splitting stage treats the original overall dataset as a cluster class, and then places the samples in the appropriate splitting positions according to the splitting strategy to obtain different subclasses. Define a dataset $D=\{P_1,...,P_i,...,P_n\}$, which contains n samples, and the ith sample $P_i=(p_1,...,p_d)$ denotes a vector with d attribute values, and classify the dataset D to obtain the classification $C=\{C_1,...,C_k\}(C_1 \cup C_2,..., \cup C_k = DandC_i \cap C_j = \varnothing, i \neq j)$. The representation of the splitting process of the dataset D at the splitting position (i, h) is shown in Equation (1) and (2).

$$C_1=\{P_j \mid P_j \in D \wedge p_i^j \leq h, i=1\sim d, j=1\sim n\} \tag{1}$$

$$C_2=\{P_j \mid P_j \in D \wedge p_i^j \leq h, i=1\sim d, j=1\sim n\} \tag{2}$$

The splitting process is iteratively processed on the dataset by using Eq. (1) and (2) until no splitting position satisfying the conditions is produced. Multiple subclasses can be obtained after processing in the splitting phase, and the merging phase requires merging these subclasses with consistent attributes. In the previous splitting process, the samples are labeled with label, and the initial value of label is 1, which is used to determine whether the subclasses in a certain level are split from the same data set. At the same time the split level marker level is added, with an initial value of 1, to update the marker for sample splitting. The splitting process is top-down, and the merging process is bottom-up, starting from the current marker until the marker is 0.

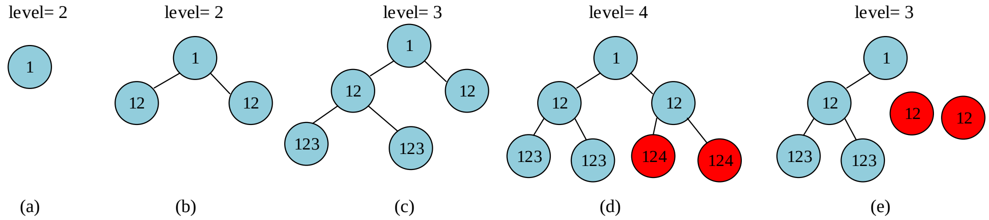


Fig. 2. Schematic diagram of the merging process.

The schematic diagram of the merging process is shown in Fig. 2, and we need to determine whether the red part in Fig. 2(d) needs to be merged. If it can be merged, then the marker level=4 is reduced by one level, that is, level=3, see Fig. 2 (a) to (d). If the condition is not satisfied, the child node of the red part replaces its parent node, and its marker level is updated and reduced by one layer, and whether to merge with other data sets is considered in the subsequent merging process, see Fig. 2(e). The condition for merging need to satisfy that the subclasses are not disconnected in any dimension and that the similarity within classes is increased and the similarity between classes is decreased after the merging process. In this study, the Calinski -Harabazs index was introduced as a measure of inter- and intra-class similarity, as shown in Eq. (3).

$$CH-index(C)=(BCSS(C)/WCSS(C))*((n-k)/(k-1)) \quad (3)$$

WCSS denotes Within Cluster Sum of Squares in Equation (3), BCSS denotes Between Cluster Sum of Squares. n and k denote the number of samples and the number of groups to be grouped, respectively. For the two subclasses C_i and C_j, their Calinski -Harabazs index after merging is calculated in Eq. (4).

$$CH-index(C_{ij})=(BCSS(C_{ij})/WCSS(C_{ij}))*((n-k_1)/(k_1-1)) \quad (4)$$

B. A Study of Hybrid Clustering Methods based on Improved Hierarchical Clustering and Density Clustering in Kissmetric Model

Hierarchical clustering-based analysis methods are parameter-insensitive and can be used for arbitrary shape class discovery, but they are computationally intensive. Density clustering-based methods can classify the original data set more rationally, but have the problem of parameter sensitivity. Based on the characteristics of both hierarchical and density clustering methods, a hybrid algorithm is combined in this study to compensate for the deficiencies and take advantage of the advantages of both hierarchical and density clustering methods. The density clustering algorithm needs to divide the high-density region from the low-density region surrounded by the fast density peaking algorithm needs to carry out the calculation of sample local density and sample distance. Eq. (5) shows the calculation of local density.

$$\rho_i=\sum_j \chi(d_{ij}-d_c),\begin{cases}\chi(x)=1, x<0\\ \chi(x)=0, x\geq 0\end{cases} \quad (5)$$

d_c denotes the truncation distance in Equation (5). The sample distances are calculated in Equation (6).

$$\delta_i=\begin{cases}\max_j(d_{ij})=1, \rho_i=\max(\rho_1,...,\rho_n)\\ \min\limits_{j:\rho_j>\rho_i}(d_{ij})=1, \rho_i\neq\max(\rho_1,...,\rho_n)\end{cases} \quad (6)$$

Based on the calculation of sample local density and sample distance, the fast density peaking algorithm is able to find the sample with high density as the center of clustering, and this sample is far away from other samples with high density, and Figure 3 shows the schematic diagram of this method.

Samples 1 and 10 in Fig. 3 represent the two clustering centers of the fast density peaking algorithm, and the method can be more successful in finding samples that can serve as clustering centers. However, there are also problems such as wrong selection of clustering centers, or selecting multiple centers in the same class and finally dividing to form multiple subclasses. In this experiment, the method is improved by first selecting multiple samples as clustering centers in the first stage of clustering using the fast density peaking algorithm, and then combining the clustering results in the subsequent hierarchical clustering. The method of cluster center selection is divided into two steps, first calculating the product of the sample distance and local density of sample i, i.e., $\gamma_i=\rho_i\times\delta_i$. Then the calculated γ_i is sorted in the order from largest to smallest, and the cluster center of the first stage selects the sample with the largest change in the value of γ. The correct clustering centers can be obtained after fast density peaking algorithm selection, and then these clustering centers need to be merged using hierarchical clustering method. In the current study the similarity measure of hierarchical clustering mainly uses average connection, full connection or single connection, but these methods do not consider the distribution of samples, which leads to the failure of the sample measure with special distribution. In this experiment, an aggregation-based hierarchical clustering method is proposed, in which a new noise point avoidance strategy is introduced to circumvent the drawbacks of artificially determining noise points and parameters. Assuming that the probability density function of subclass C_i is $f_i(v)$, the probability density function of subclass C_j is $f_j(v)$, and v denotes the attribute values of the samples within the class, the definitions of the connectivity of C_i and C_j in Equation (7) can be obtained.

$$join(C_i,C_j)=\sum_{p\in C_i\cup C_j}\min(f_i(p),f_j(p)) \quad (7)$$

This leads to the aggregation function of C_i and C_j in Eq. (8).

$$cohesion(C_i,C_j)=\frac{join(C_i,C_j)}{|C_i|+|C_j|} \quad (8)$$

$|C_i|$ denotes the number of samples in C_i and $|C_j|$ denotes the number of samples in C_j in Equation (8). Assuming that the samples in the subclass C_i obey a multivariate normal distribution, i.e., $V\sim N_d(\mu,\psi)$, the expression of the probability density function of C_i is given in Eq. (9).

$$f_i(v)=(2\pi)^{-\frac{d}{2}}(\det\psi)^{-\frac{1}{2}}\exp[-\frac{1}{2}(v-\mu)^T\psi^{-1}(v-\mu)] \quad (9)$$

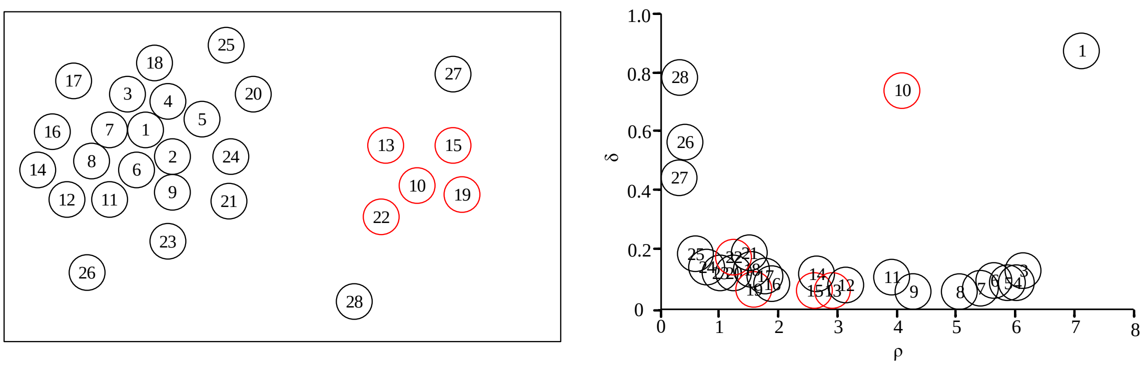


Fig. 3. Schematic diagram of the fast density peaking algorithm.

d denotes the dimensionality, μ denotes the mean vector, and ψ denotes the covariance matrix in Eq. (9). The calculation of μ is shown in Eq. (10).

$$\mu = \frac{1}{n}\sum_{i=1}^{n} v_i \tag{10}$$

The calculation of ψ is shown in Eq. (11).

$$\psi = \frac{1}{n}\sum_{i=1}^{n}(v_i-\mu)(v_i-\mu)^T \tag{11}$$

Given the probability density function of C_i according to Eq. (9), the volume calculation of C_i can be obtained from this, see Eq. (12).

$$V_i = \frac{4}{3}\pi\sqrt{\lambda_1\lambda_2,...,\lambda_d} \tag{12}$$

$\lambda_1\lambda_2,...,\lambda_d$ represents the d eigenvalues of the covariance matrix ψ in Eq. (12). Because the subclass volume and the eigenvalues of ψ are proportional, the constant term is removed for the subsequent calculation, i.e., $V_i = \sqrt{\lambda_1\lambda_2,...,\lambda_d}$. Eq. (13) is the density calculation of C_i.

$$D_i = \frac{|C_i|}{V_i} \tag{13}$$

According to the characteristics of noise points, if two subclasses that are close to each other have a large difference in density or volume, one of them has a higher probability of being a noise point cluster. Based on this feature, the definition of the noise point avoidance function for the subclasses C_i and C_j is given in Eq. (14).

$$dependence(C_i,C_j) = \frac{D_i+D_j}{2\times\sqrt{D_i\times D_j}} + \frac{V_i+V_j}{2\times\sqrt{V_i\times V_j}} + \frac{\frac{1}{2\times d}\sum_{t=1}^{d}(\sqrt{\lambda_t^i}+\sqrt{\lambda_t^j}}{\left|\overline{c_i}-\overline{c_j}\right|} \tag{14}$$

$\overline{c_i}$ denotes the cluster center of subclass C_i in Equation (14), $\overline{c_j}$ denotes the cluster center of subclass C_j, $\sqrt{\lambda_t^i}$ denotes the length of each dimensional axis of C_i, and $\sqrt{\lambda_t^j}$ denotes the length of each dimensional axis of C_j. Since $\frac{D_i+D_j}{2\times\sqrt{D_i\times D_j}}\le 1, \frac{V_i+V_j}{2\times\sqrt{V_i\times V_j}}\le 1$, the value of the mean inequality is maximized when and only when $D_i = D_j, C_i = C_j$. From the noise point avoidance function in Formula (14), the closer the density and volume of the two subcategories are, the higher the dependency of the two subcategories is, and the greater the probability of their combination is. On the contrary, the combination probability of the two is smaller. According to the definitions of noise point avoidance function and aggregation function, the similarity of subclasses C_i and C_j is given in Eq. (15).

$$similarity(C_i,C_j) = cohesion(C_i,C_j) + dependence(C_i,C_j) \tag{15}$$

The similarity measure in this experiment consists of an aggregation function and a noise point processing function, which fully considers the distribution of samples within classes while circumventing the interference of noise points during class merging.

Kissmetric is an automated customer engagement data analysis model based on the hierarchical clustering approach described above, capable of providing business data analysis to clients. The main components of the Kissmetric data analysis model include the visitors to the web pages, the features used to describe the user information, and the events and their attributes. The hierarchical clustering method is the key method used in the Kissmetric data analysis model for statistical analysis of data and is divided into three main steps. The first step is to count and divide the number of customers, which can get three types of customers, namely, resource customers who have browsed the products, potential customers and customers who have purchased the products. The second step is to count and divide the types of goods, mainly by analyzing the information of the type, name, price and size of the goods. The third step is to count and divide the platforms, and make a regional division of the number of customers who have purchased goods on different platforms.

IV. Performance Study of Business Data Analysis Model based on Kissmetric

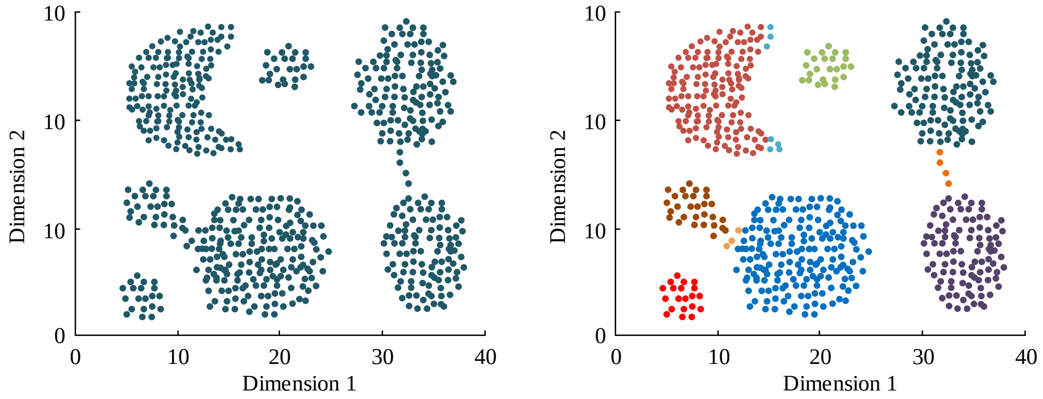
The parameter settings in the experiment are as follows: the number of centroids is 8, and the maximum number of iterations is 300. In the improved hierarchical clustering method, the effectiveness of the algorithm in the splitting process is verified using the Aggregation dataset in Fig. 4. For the original Aggregation dataset, the improved hierarchical clustering method is able to split the original samples and obtain multiple subclasses.

The original data set needs to be merged after the splitting process. In Fig. 5, the subclasses with consistent attributes need to be merged in the merging stage to finally obtain a more accurate classification effect, and the clustering accuracy of this method is 99.21%.

Performance validation for clustering algorithms can be evaluated using the F-measure metric, which is a weighted summed average of recall and accuracy, and is used to evaluate the merit of the classification model. The Rand Index (RI) can be used to measure the similarity of clustering results, but there is a lack of differentiation. To address this problem, the Adjusted Rand Index (ARI) makes some improvements on the basis of RI, which can make a clearer distinction of clustering effects. In the range of [-1,1], the larger value of ARI indicates the better effect of the clustering method. The results of the F-measure metrics and ARI of the hybrid clustering method proposed in this experiment compared with K-means, K-medoids, and K-means++ methods in different dimensions in Fig. 6 [19-21]. The highest F-measure index and ARI values of the hybrid clustering method proposed in this experiment are 0.997 and 0.998, respectively, under different dimensions, which are higher than those of K-means, K-medoids, and K-means++ methods.

The F-measure and ARI values of the algorithm are compared in Fig. 7 for different numbers of classes of the dataset. The highest F-measure metrics and ARI values of the hybrid clustering method proposed in this experiment are 1.000 and 0.999, respectively, under different numbers of data set classes, which are higher than those of K-means, K-medoids, and K-means++ methods. F-measure index and ARI values of the hybrid clustering method do not change due to the number of classes in the data set, and its clustering effect is better and more stable.

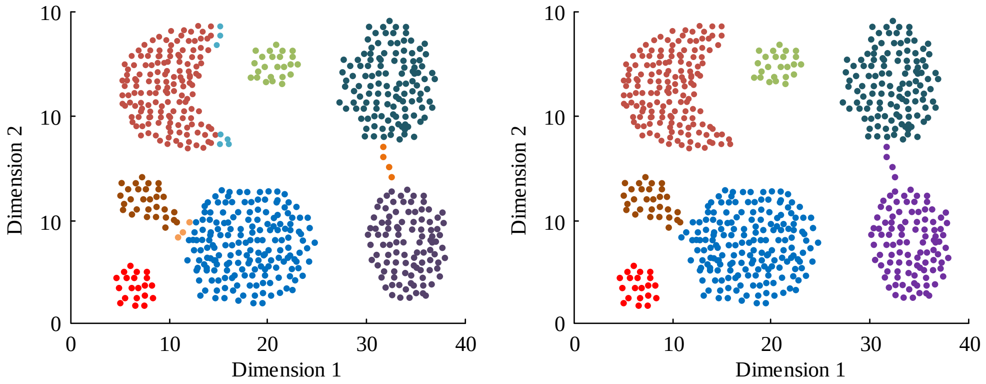
The hybrid clustering method proposed in this experiment is used with K-means, K-medoids, and K-means++ methods to cluster the Aggregation dataset in Fig. 8. The hybrid clustering method accurately divides the Aggregation dataset into seven classes according to the similarity calculation results, while the rest of the methods for some of the samples were classified with ambiguity, and more than seven classes were finally classified. The hybrid clustering method proposed in this experiment has a better clustering effect and its clustering accuracy is better.



(a) Dataset Aggregation

(b) Aggregation Split Result

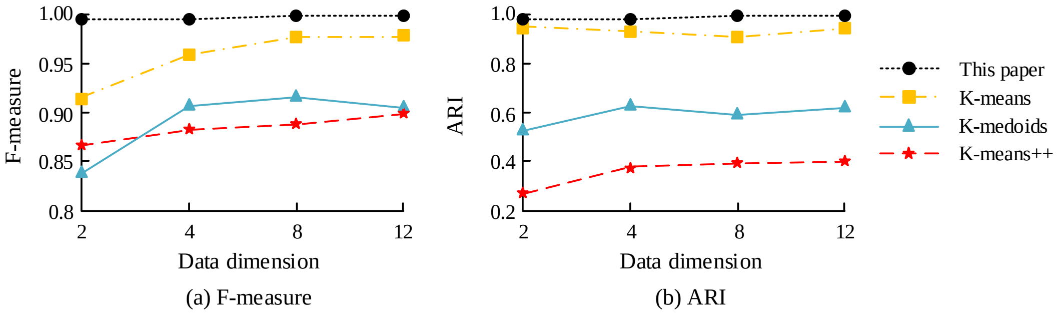
Fig. 4. Cracking effect of improved hierarchical clustering method.



(a) Aggregation Split Result

(b) Aggregation Merge Result

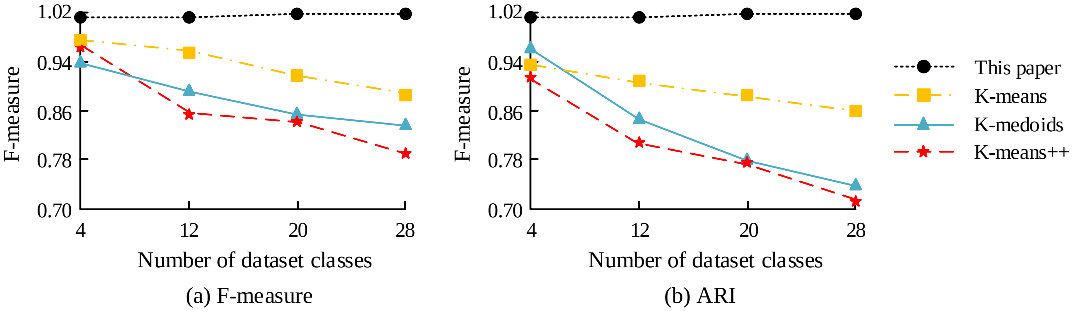
Fig. 5. Merging effect of improved hierarchical clustering method.



(a) F-measure

(b) ARI

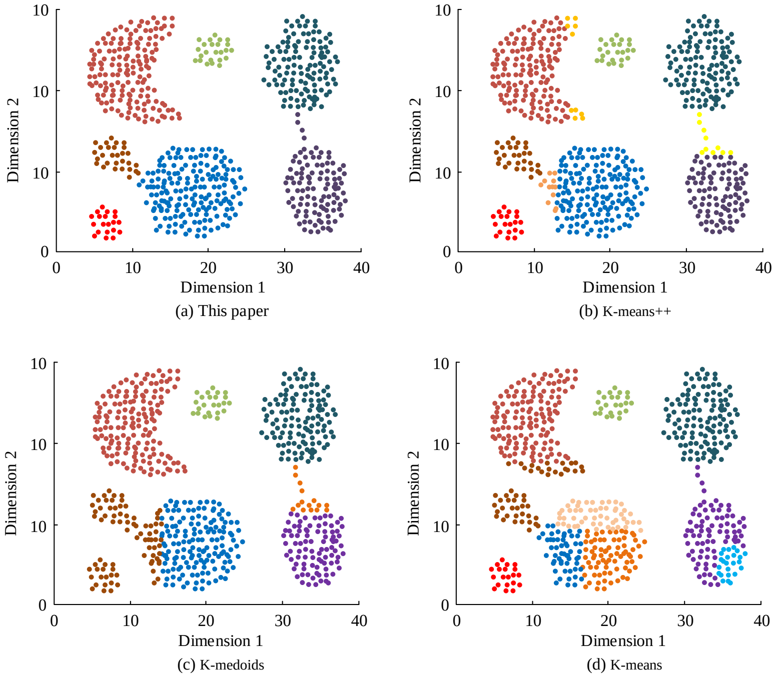
Fig. 6. F-measure and ARI values of the algorithm in different dimensions.



(a) F-measure

(b) ARI

Fig. 7. F-measure and ARI values of the algorithm for different number of classes of the dataset.



(a) This paper

(b) K-means++

(c) K-medoids

(d) K-means

Fig. 8. Comparison of clustering effects of different algorithms.

To further validate the performance of the hybrid clustering method, experiments were conducted to evaluate the performance using Purity validity metrics in eight datasets, which included simulated datasets 1-6 and real datasets 7-8. Table I shows the results in simulated datasets 1-6. When running datasets 1-6, the mean accuracy and mean squared error of the hybrid clustering method were the best among the four algorithms, which were 95.94% vs. 5.89%, 94.72% vs. 0.57%, 89.72% vs. 4.97%, 87.45% vs. 5.53%, 93.83% vs. 5.76%, and 88.43% vs. 5.40%, indicating that the algorithm has higher accuracy. When comparing the clustering analysis of datasets 1 to 6 with different methods, the hybrid clustering method has the highest clustering accuracy, the best clustering effect and the best stability in the six datasets. The hybrid clustering method has significantly improved in terms of accuracy and stability when clustering the simulated datasets. In terms of algorithm runtime, the hybrid clustering method has improved over the other methods.

TABLE I. EXPERIMENTAL RESULTS OF SIX SIMULATED DATA SETS

Algorithm	Parameter	Data set 1		Data set 2		Data set 3		Data set 4		Data set 5		Data set 6	
		Mean value	Mean square value	Mean value	Mean square value	Mean value	Mean square value	Mean value	Mean square value	Mean value	Mean square value	Mean value	Mean square value
This paper	P(%)	95.94	5.89	94.72	0.57	89.72	4.97	87.45	5.53	93.83	5.76	88.43	5.40
	Time consuming(ms)	52.42	6.05	372.96	68.54	256.03	22.18	387.07	21.17	46.37	12.10	134.06	24.19
K-means++	P(%)	93.24	6.55	93.93	1.87	83.41	8.01	85.17	8.71	90.94	8.86	84.19	6.58
	Time consuming(ms)	48.38	18.14	158.26	27.22	163.30	33.26	186.48	51.41	32.26	13.10	108.86	30.24
K-medoids	P(%)	92.62	8.72	87.19	6.69	83.07	6.15	85.34	5.61	85.84	12.92	78.50	9.15
	Time consuming(ms)	25.20	4.03	74.59	15.12	79.63	10.08	92.74	18.14	19.15	6.05	47.38	11.09
K-means	P(%)	91.98	6.38	93.10	2.78	86.07	8.66	85.56	5.86	86.15	6.97	83.43	5.52
	Time consuming(ms)	56.45	14.11	192.53	49.39	131.04	25.20	201.60	51.41	31.25	12.10	102.82	24.19

TABLE II. EXPERIMENTAL RESULTS OF TWO REAL DATA SETS

Algorithm	Parameter	Data set 7		Data set 8	
		Mean value	Mean square value	Mean value	Mean square value
This paper	P(%)	89.71	6.17	88.85	0.33
	Time consuming(ms)	28.22	6.05	24.19	2.02
K-means++	P(%)	81.32	10.77	88.56	0.34
	Time consuming(ms)	24.19	6.05	18.14	3.02
K-medoids	P(%)	86.15	11.29	85.20	7.65
	Time consuming(ms)	9.07	3.02	17.14	2.02
K-means	P(%)	85.48	9.35	82.03	10.45
	Time consuming(ms)	20.16	4.03	35.28	8.06

The results obtained in the real dataset are presented in Table II. The hybrid clustering method has the best quality and stability of clustering results with the best accuracy mean and accuracy mean squared error of 89.71 % vs. 6.17 % and 88.85 % vs. 0.33 %, respectively, when dealing with datasets 7 and 8. When comparing the clustering analysis of the real datasets 7 and 8 with different methods, the hybrid clustering method has the highest clustering accuracy, the best clustering effect and the best stability in these two datasets. The results indicate that the hybrid clustering method has significantly improved in terms of accuracy and stability when clustering analysis is performed on the real dataset. The results from the real and simulated datasets mentioned above show that the algorithm proposed in this experiment can handle datasets of any shape and different distributions. And this algorithm has higher accuracy compared to most current clustering algorithms. This is because the algorithm proposed in this experiment can avoid the problem of missed or incorrect selection when selecting cluster centers. Therefore, this algorithm can correctly classify samples. At the same time, this method introduces a similarity measurement method. It can be used to process datasets with different types and uneven sample distribution. At the same time, the treatment of noise points was added in the experiment. This can effectively handle the impact of noise points on clustering results.

V. RESULTS AND CONCLUSIONS

The improvement of data mining technology promotes the efficiency of commercial data application. In existing research, the K-means clustering algorithm can effectively handle data of different scales. In order to improve the ability of traditional clustering algorithms to determine the cluster center, some scholars proposed a hierarchical clustering analysis method [8]. This method can reduce the subjective factors affecting the determination of k-values in traditional clustering methods. However, this method requires a large amount of computation and cannot trace back to the intermediate clustering process. For this reason, an improved hierarchical clustering algorithm is proposed in the experiment, which is generated in the split phase and the merge phase. This algorithm can combine the algorithm with density clustering methods while considering noise point processing, achieving automatic determination of clustering centers and improving clustering performance. For the original Aggregation dataset, the improved hierarchical clustering method can crack the original samples to obtain multiple subclasses. Then these subclasses with consistent attributes are merged in the merging stage to get more accurate classification results, and the clustering accuracy of this method is 99.21%. Under different dimensions, the highest F-measure index and ARI values of the hybrid clustering method proposed in this experiment are 0.997 and 0.998, respectively. Under different numbers of classes in the data set, the highest F-measure index and ARI values of the hybrid clustering method are 1.000 and 0.999, respectively, which are higher than those of the K-means, K-medoids, and K-means++ methods. When running simulated datasets 1 to 6, the mean and mean squared error of the hybrid clustering method were the best among the four algorithms, with 95.94% vs. 5.89%, 94.72% vs. 0.57%, 89.72% vs. 4.97%, 87.45% vs. 5.53%, 93.83% vs. 5.76%, and 88.43% vs. 5.40 %, respectively, indicating that the algorithm has higher accuracy. When dealing with the real data sets 7 and 8, the hybrid clustering method has the best accuracy mean, and accuracy mean squared error of 89.71% versus 6.17% and 88.85% versus 0.33%, respectively. From the results, F-measure index and ARI values of the hybrid clustering method do not change due to the influence of dimensionality and the number of classes in the dataset, and the quality and stability of its clustering results are better. The validation results in different datasets show that the method established in this experiment can handle datasets of any shape and different distributions. And this algorithm has higher accuracy compared to most current clustering algorithms. Compared to the K-means, K-medoids, and K-means++methods in references [19-21], its F-measure, ARI indicators, accuracy, and other indicators are higher, and they have certain advantages. This is because the algorithm proposed in this experiment can avoid the problem of missed or incorrect selection when selecting cluster centers. Therefore, this algorithm can correctly classify samples. At the same time, this method introduces a similarity measurement method. It can be used to process datasets with different types and uneven sample distribution. At the same time, the treatment of noise points was added in the experiment. This can effectively handle the impact of noise points on clustering results. Although the method proposed in this experiment can correctly and effectively handle different types of data, there are still some shortcomings in this method. As the amount of data increases, the clustering performance of the method will decrease. The dataset used in this experiment has a smaller scale. The application effect of the improved method in large-scale data is still uncertain. Therefore, improving the ability of clustering algorithms to handle massive amounts of data is an important task in the subsequent work of this article. In addition, the data in real life is quite complex and noisy. This will affect the stability of the clustering algorithm. Therefore, how to effectively improve the anti-interference ability and practicality of clustering algorithms, as well as further enhance the algorithm's ability to process real data, is one of the next directions that need to be studied.

References

[1] Tao D, Yang P, Feng H. Utilization of text mining as a big data analysis tool for food science and nutrition. Comprehensive Reviews in Food Science and Food Safety, 2020, 19(2): 875 - 894.

[2] Alim A, Shukla D. Sampling - based estimation method for parameter estimation in big data business era. Journal of Advances in Management Research, 2020, 18(2): 297 - 322.

[3] Cope J M, Gertseva V. A new way to visualize and report structural and data uncertainty in stock assessments. Canadian Journal of Fisheries and Aquatic Sciences, 2020, 77(8): 1275 - 1280.

[4] Jiang L. A Study on the Application of Statistical Analysis Method of Big Data in Economic Management. Journal of Commercial Economics, 2020, 3(3): 69 - 72.

[5] Ciobotaru G, Chankov S. Towards a taxonomy of crowdsourced delivery business models. International Journal of Physical Distribution & Logistics Management, 2021, 51(5) : 460 - 485.

[6] Han H, Zhou M C, Shang X, et al. KISS+ for Rapid and Accurate Pedestrian Re - Identification. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(1): 394 - 403.

[7] Zeng F, Zhang W, Zhang S, Zheng N. Re - KISSME: A robust resampling scheme for distance metric learning in the presence of label noise. Neurocomputing, 2019, 330(FEB.22): 138 - 150.

[8] Hulme P E. Hierarchical cluster analysis of herbicide modes of action reveals distinct classes of multiple resistance in weeds. Pest Management Science, 2022, 78(3): 1265-1271.

[9] Arunkumar P M, Kannimuthu S. Mining big data streams using business analytics tools: a bird's eye view on MOA and SAMOA. International Journal of Business Intelligence and Data Mining, 2020, 17(2): 226 – 236.

[10] Tao D, Yang P, Feng H. Utilization of text mining as a big data analysis tool for food science and nutrition. Comprehensive Reviews in Food Science and Food Safety, 2020, 19(2): 875 - 894.

[11] WA Xi, ZW B, MSC D, LA Qi, WS A. An adaptive and opposite K - means operation based memetic algorithm for data clustering. Neurocomputing, 2021, 437(8): 131 - 142.

[12] S Liu, Q Lin, KC Wong, CAC Coello, J Zhang. A Self - Guided Reference Vector Strategy for Many - Objective Optimization. IEEE Transactions on Cybernetics, 2020, 52(2): 1164 - 1178.

[13] Molokomme D N, Chabalala C S, Bokoro P N. Enhancement of Advanced Metering Infrastructure Performance Using Unsupervised K - Means Clustering Algorithm. Energies, 2021, 14(9): 2732 - 2743.

[14] X Qin, J Li, W Hu, J Yang. Machine Learning K - Means Clustering Algorithm for Interpolative Separable Density Fitting to Accelerate Hybrid Functional Calculations with Numerical Atomic Orbitals. The Journal of Physical Chemistry A, 2020, 124(48): 10066 - 10074.

[15] Dressler M, I Paunovíc. Business Model Innovation: Strategic Expansion of German Small and Medium Wineries into Hospitality and Tourism. Administrative Sciences, 2021, 11(4): 146 - 157.

[16] Reyes - Ruiz G, M Hernández - Hernández. Fuzzy clustering as a new grouping technique to define the business size of SMEs through their financial information. Journal of Intelligent and Fuzzy Systems, 2021, 40(2): 1773 - 1782.

[17] Gubu L, Rosadi D, Abdurakhman A. Robust Portfolio Selection with Clustering Based on B usiness Sector of Stocks . Media Statistika, 2021, 14(1): 33 - 43.

[18] Liashenko O, Kravets T, Prokopenko M. Consumer behavior clustering of food retail chains by machine learning algorithms. Access Journal, 2021, 2(3): 234 - 251.

[19] Sailekha K, Deluxni N. Comparative Analysis of Customer Behaviour using K - means Algorithm Over Convolutional Neural Network with Increase Inaccuracy of Prediction. ECS transactions, 2022, 107(1): 12459 - 12471.

[20] Lund B, Ma J. A review of cluster analysis techniques and their uses in library and information science research: k - means and k - medoids clustering. Performance measurement and metrics: The international journal for library and information services, 2021, 22(3): 161 - 173.

[21] Xiong Z, Li J, Wu H, et al. Understanding operation patterns of urban online ride - hailing services: A case study of Xiamen. 2021, 101(C): 100 - 118.