Customer Market Analysis Based on Interval Value Data Dynamic Clustering Algorithm

Yihan Hu
University College London
London, United Kingdom
1757700544qq.com

2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS) | 979-8-3503-1545-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIICS59993.2023.10421290

***Abstract*—The traditional K-means clustering center initial value algorithm does not perform well in massive data. In order to improve the initialization efficiency of the clustering center, this article adopted the idea of dynamic clustering algorithm for inter region value data to classify the data, and selected the center of dense grids as the initial clustering center based on the distribution characteristics of the data. This method can ensure that the cluster center is located in a high-density area rather than being overly concentrated, allowing the K-Means algorithm to quickly and efficiently initialize the cluster center when processing massive data. Finally, this article compared and analyzed the performance of data clustering methods using traditional K-means algorithm and K-means++algorithm through simulation experiments. The evaluation indicators of the experiment are CR index, Silhouette Score, and accuracy. From the comprehensive analysis of three indicators, it can be seen that the performance of customer market analysis data based on this algorithm was superior to other algorithms (Silhouette Score: this algorithm was 0.1249 higher than K-means++and 0.4903 higher than K-means algorithm). The results indicated that compared with traditional clustering methods, clustering methods had significant advantages in the clustering process.**

Keywords—customer market analysis, interval value data dynamics, clustering algorithm, K-means algorithm

I. INTRODUCTION

The traditional "user analysis" technology represented by the RFM (Recency Frequency Monthly) model is limited in processing complex data structures and their relationships due to its dependence on feature selection. A large number of studies have applied data mining technology to customer analysis and operational decision-making assistance, which not only reflects the high value of clustering applications in data mining environments, but also reflects the hope of dividing the same object into multiple clusters in applications with similar customer segmentation goals, so that the objects in each cluster have high similarity, thereby better customer segmentation. Among them, the K-means method, due to its good scalability and high computational efficiency, is particularly suitable for large-scale data and has become a classic clustering algorithm.

Experts have conducted relevant research on the analysis methods for current customer market classification. Abdulhafedh A explored a problem of using cluster algorithms to determine the market strategy of a credit card company. Customer classification refers to dividing customers into several groups based on their common characteristics, which is very helpful for banks and enterprises. K-means clustering, hierarchical clustering, and PCA (principal component analysis) were explored to determine the customer base of enterprises. His research selected 8950 credit card consumption data within six months and used clustering analysis methods to accurately classify them. He conducted research from two aspects: firstly, he adopted hierarchical clustering and K-means clustering, taking into account various factors comprehensively. Secondly, the PCA method was used to reduce the dimensionality of the samples. He selected the optimal number of clusters and then used the updated number of clusters to re cluster the samples. The results showed that principal component analysis is an effective method that can be used to detect K-means and hierarchical clustering [1]. Bandyopadhyay S introduced a computer-based intelligent technology that can help customers meet their shopping requirements. In addition, it can also assist retailers in supply chain management and develop corresponding business strategies based on current market conditions. In order to gain a competitive advantage in the market and maximize customer value, he used the K-means clustering method to divide the customer group into different categories, thereby obtaining different types of customer groups. The use of principal component analysis can effectively reduce the dimensions of products and customers. Bandyopadhyay S focused on identifying potential differences in brand, product, and price perception among customers through their shopping habits. The results showed that the clustering results obtained based on PCA and K-Means algorithms were very similar. From the feedback of existing customers, the results were acceptable, and the customer's needs were met according to the quantity (price range) they wanted to consume during online shopping [2]. By using K-means and applied science, Anitha P analyzed a set of real-time trade and retail industry data. The values and parameters in the dataset are assigned over a specific time period, which can help people better understand the shopping methods and behaviors of customers in different regions. On this basis, the K-Mean algorithm was used to partition the samples. By calculating the contour coefficient, different types of data clusters were tested. The obtained sales and transaction results were compared with different parameters such as sales cycle, sales frequency, and sales volume [3]. Chaudhary K uses the concept of big data technology to process and analyze data and predict consumer behavior on social media. He analyzed the consumer behavior on social media platforms based on various parameters and criteria. He analyzed the perception and attitude of consumers towards social media platforms. To get high quality results, he preprocessed the data using different data preprocessing methods to detect outliers, noise, errors and duplicate records. Using machine learning, he created a mathematical model to predict consumer behavior on social media platforms. The model is a predictive model for predicting consumer behavior on social media platforms. 80% of the data was used for training and 20% for testing and the experiments verified the effectiveness of the model [4]. There are fewer studies on how to utilize data for value creation in enterprises with massive amounts of data (hundreds of transactions and tens of attributes). Based on the theory of "Big Data Value Creation", Hopf K conducted an

empirical study on a "resource-based" renewable energy retailer. Companies in this industry have high operating costs but low sales revenues, and therefore limited access to data. His findings reveal that data analysis and value generation mechanisms (democratization, contextualization, data experimentation, and data insights) are also effective under data-constrained conditions. Therefore, in practical applications, it is important to focus on the type and amount of data as well as the contextual factors of managers (clear strategy, vision, leadership) and all employees (including openness to agile working models, data awareness, etc.) [5]. While the field of management has recognized the importance of using big data to understand, predict, and respond to future emergencies, there is currently a lack of scientific understanding of these issues, especially in the face of the daunting challenges of the new Crown Pneumonia outbreak.Sheng J's presentation will discuss in detail the "black swan" phenomena of description/diagnosis, prediction, and specification, and how they are being used to study global crises such as the COVID-19 pandemic and their implications for operators and policy makers [6].Although the above research can achieve effective analysis of customer consumption, existing clustering methods still have many shortcomings, such as uncertain number of clusters, sensitivity to initial cluster centers and individual cluster points, low efficiency in direct use, and the large computational complexity of most methods, making it difficult to apply in big data environments.

This article discussed customer market analysis based on interval value data dynamic clustering algorithm. The improved clustering algorithm can fill the shortcomings of existing clustering algorithms, which greatly improves its efficiency. Moreover, interval value data can better match the uncertainty of customers' shopping behavior in reality compared to massive data, which is more conducive to data analysis and has research significance.

II. K-means Clustering Algorithm and Customer Market Analysis

A. K-means Clustering Algorithm

The K-means based clustering method first selects K data objects from the original dataset as the initial centers for clustering. The algorithm takes K selected targets from the original dataset as the initial clustering center, and divides the target into corresponding classifications through Euclidean distance, thereby calculating the clustering center of the classification. Then, Euclidean distance is used to attribute the samples to the corresponding classification, and the classification of the new classification is calculated until the convergence of the classification was achieved [7-8]. Usually, the evaluation function can be represented by the sum of squares of intra cluster errors: $E = \sum_{i=1}^{k} \sum_{p \in c_i} dis(p, m_i)^2$.

Among them, k represents the number of clusters; p represents the data object in cluster i; m_i represents the cluster center, and the cluster center updates itself according to the following equation: $m_i = \frac{1}{|m_i|} \sum_{i=1}^{|m_i|} x_i = \frac{1}{|m_i|} \left(x_1 + x_2 + \cdots + x_{|m_i|} \right)$.

$|m_i|$ refers to the number of data objects [9].

B. Practical Significance of Customer Market Analysis

With the intensification of market competition in various aviation industries, customer value has become increasingly important. Therefore, accurate and effective analysis of customer value has become a key factor in company decision-making [10]. The fundamental value of customer loyalty is a very important part of a company's strategy: only by accurately measuring customer value and continuously increasing customer dependence on the company can the company gain customer loyalty and maintain a competitive advantage at all times [11]. In practical work, through the analysis of multiple levels of customer value demand and supply and demand, it is possible to understand what customers are most concerned about and how their interests change over time [12]. Afterwards, in every step of delivering value to customers, a comprehensive customer value measurement system is established, dividing customers with different values into categories and providing targeted customer service to customers with different values, thereby maintaining customers' sustained trust in the company and increasing its stickiness [13]. For example, the customers of airlines are generally high-end customers, and the working-class are more inclined to choose aviation as the mode of transportation. Therefore, the results of customer market analysis based on interval value data dynamic clustering algorithms can provide technical support for most airlines to conduct customer value analysis and help them find suitable high-value customers, thus providing convenience for them. At the same time, it also contributes to the development of the overall aviation industry [14].

III. Customer Market Segmentation Based on Interval Value Data Dynamic Clustering

A. Data Preprocessing

With the advent of the information age, humanity is facing an increasing amount of massive data, and traditional analysis methods are no longer able to handle massive data well. In the late 1980s, Europe developed a new data processing method - Signal data analysis (SDA). Contextual perception refers to analyzing the information of object terminals to determine the correlation between elements and thus determine the structure between elements [15]. Due to data errors caused by measurement, calculation, and other factors, coupled with data loss caused by incomplete data, the input object data is often an uncertain numerical value. Therefore, using intervals to express these uncertain object data is more in line with people's ideas and closer to reality than expressing real object data.

In order to convert conventional information into interval value information, it is necessary to convert it into interval value information. In order to better record customers' purchasing behavior, the company has established a transaction database, assuming that each record is organized according to (customer ID, product ID, product name, transaction quantity, transaction price, transaction time) [16]. Due to the influence of personal preferences, economic conditions, and other factors on consumers' purchasing behavior, in real purchasing activities, consumers' purchasing behavior has strong randomness and uncertainty. The preprocessing of data is based on customer ID, which is calculated and summarized according to the attributes that best reflect the characteristics of customer purchasing behavior during a specific period. Then, the traditional

transaction record value is converted into interval value data [17]. On the basis of preprocessing the data, a target customer representation method based on multidimensional interval value vectors is proposed, as shown in Fig. 1.

Customer ID	Product ID	Product Name	Transaction Quantity	Price	Empty time
Product ID	121	Clothing	6	1200	2020.10.08
Product Name	122	Skincare	7	830	2020.10.30
Transaction Quantity	123	Purses	1	60	2020.11.10
Price	124	Watches	2	540	2020.10.28
Empty time	……	……	……	……	……

Customer ID	Quantity	Price	Frequency
30001	[6,28]	[1000,2000]	[6,9]
30002	[20,30]	[670,1000]	[9,25]
30003	[7,10]	[70,200]	[5,13]
……	……	……	……

Fig. 1. Interval Values from Original Transaction Records to Data Preprocessing

B. *Adaptive Euclidean Distance*

If there are two sets of interval functions $X_i(z)=(x_i^m(z),x_i^n(z))$ and $X_j(z)=(x_j^m(z),x_j^n(z))$, then the midpoint function is $x_i^c(z)=\dfrac{\left|x_i^m(z)+x_j^n(z)\right|}{2}$ and the radius function is $x_i^r(z)=\dfrac{\left|x_i^m(z)-x_i^n(z)\right|}{2}$. Similarly, $x_j^c(z)$ and $x_j^r(z)$ can be obtained. The Euclidean distance between interval functions $X_i(z)$ and $X_j(z)$ can be expressed as:

$$d_k=\sqrt{\int\left(x_i^m(z)-x_j^m(z)\right)^2+\left(x_i^n(z)-x_j^n(z)\right)^2dz}$$
$$=2\int\left(x_i^c(z)-x_j^c(z)\right)^2+\left(x_i^r(z)-x_j^r(z)\right)^2dz \quad (1)$$

Firstly, the interval function data $X_i(z)$ and $X_j(z)$ are calculated based on Formula (1) to obtain the distance between samples, and the two samples with the largest distance are selected as the initial clustering centers, denoted as $c_1(z)$ and $c_2(z)$. If the number of clusters k $\geqslant$ 3, the sample with the maximum sum of distances from the first two cluster centers would be used as the third initial cluster center $c_3(z)$, thus obtaining all k initial cluster centers: $c(z)=\{c_1(z),c_2(z),\cdots,c_k(z)\}$ [18].

C. *Class Initialization for Dynamic Clustering*

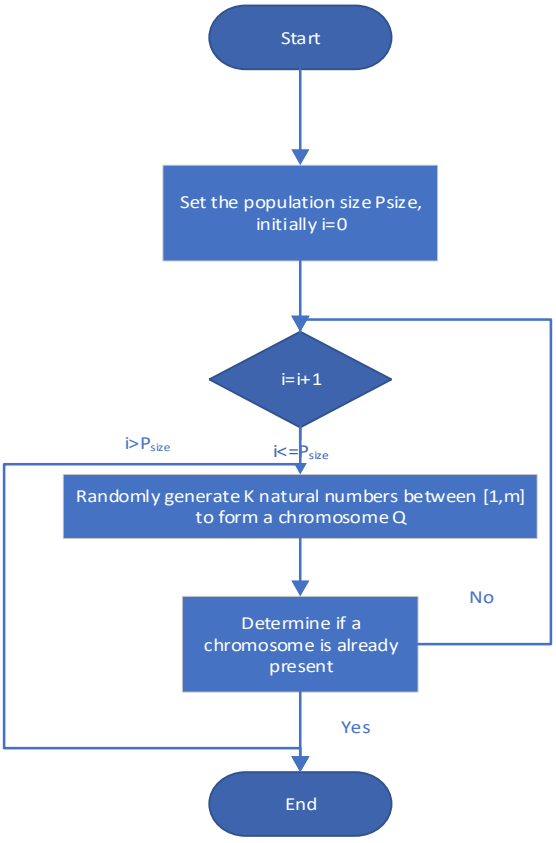

Fig. 2. Process of Genetic Algorithm

Class initialization is the first step in the entire clustering process. If the initial class initialization fails to obtain a reasonable class initialization, it would have a negative impact on the subsequent dynamic clustering process, leading to a decrease in clustering performance. The genetic algorithm is used for clustering analysis of data, and the algorithm steps are shown in Fig. 2 [19].

1) Algorithm termination conditions: If the evolutionary algebra exceeds the maximum genetic algebra Q, or if the number of iterations of the same optimal result continuously exceeds a certain threshold δ during execution, the algorithm would terminate [20-21].

2) The Entire Algorithm Process:

a) Initializing the consumer group: based on group size P_{size} *, a randomly selected customer is used as the clustering center to form a single customer, with* $C=\{c_1,c_2,\cdots,c_k\}$ *as the encoding and customer ID as the chromosome gene.*

b) Performing K-means on each individual: each data object is assigned a nearest cluster, and then wait for the assignment to end before calculating a cluster.

c) Calculate the degree of adaptation of each individual: if the child of the individual has a higher degree of adaptation than its parent, the child replaces the surrogate parent.

d) Cross talk: each individual in a population undergoes selection, inbreeding, and genetic mutations. After each evolution, the winner is replaced by the winner from two populations through cross mutation. If the adaptability of a group is stronger than that of its parents, the parents would be replaced by the group.

If the overall termination criteria are met, it is entered into step (6), and vice versa, it is entered into step (2).

In two groups, the individual with the maximum fit value is used as the initial cluster, and K-means algorithm is used to cluster and analyze the customer cluster, obtaining the final clustering results [22-23].

IV. Algorithm Simulation Based on Constructing Interval Datasets

When simulating on the basis of establishing interval arrays, it is necessary to establish them on the two-dimensional real space R. This study generated four different types of data, each containing 300 pieces of data, with 150 pieces of data and 90 pieces of data. When establishing an interval set, it is necessary to first establish a midpoint set, and then establish a circle according to the given number of intervals within the circumference range, thus forming a circle. It is assumed that the midpoint of each type of interval data is determined by two variables that conform to a normal distribution, the mean and covariance matrix of that variable can be expressed as:

$$\begin{cases} \mu=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} \\ \Sigma=\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix} \end{cases} \quad (2)$$

Each data point (x, y) serves as the center point of the interval dataset. At the same time, parameters γ_1 and γ_2 generated based on the set range are used as radii for the corresponding center point data on the x-axis and y-axis, which can be expanded into a two-dimensional interval data $([x-\gamma_1,x+\gamma_1],[x-\gamma_2,x+\gamma_2])$, with radius ranges of [1,5], [1,10], [1,15], and [1,20], respectively.

The values of the mean and variance matrix related parameters for the four types of datasets are shown in Table I.

TABLE I. Mean and Variance of the Dataset

Data sets	μ_1	μ_2	σ_1	σ_2
Class 1	28	55	121	9
Class 2	82	49	16	81
Class 3	48	35	81	16
Class 4	56	78	25	25

The random interval dataset obtained based on the parameters in Table I is shown in Fig. 3.

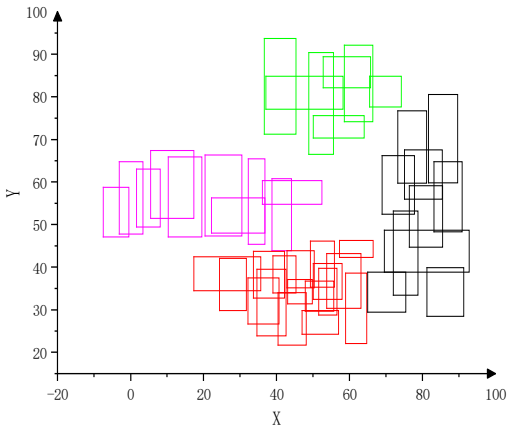

Fig. 3. Generated Interval Dataset

From Fig. 3, it can be seen that the data can be clustered into four categories. According to the above process steps of customer market segmentation based on interval value data dynamic clustering, the effect diagram of clustering results for generating interval datasets in Fig. 4 can be obtained.

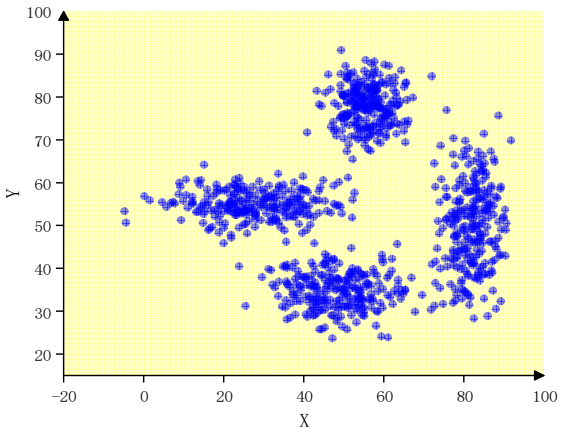

Fig. 4. Clustering Effect of Generating Interval Datasets

Fig. 4 shows the integration effect of a set of spatial data. Cluster effect map is an important method for qualitative evaluation of visual clustering results. Clustering instances

include data with multiple attributes and different dimensions. The clustering results can directly reflect the attributes and topological relationships of the original data, and can effectively map high-dimensional data to low-dimensional space. The clustering effect can be seen in Fig. 4. By classifying customer data, it is possible to accurately classify them and achieve excellent clustering results. At the same time, K-means algorithm, K-means++algorithm, and our algorithm were used for clustering comparison of data, analyzing CR index, Silhouette Score, and accuracy. The analysis results are as shown in Fig. 5.

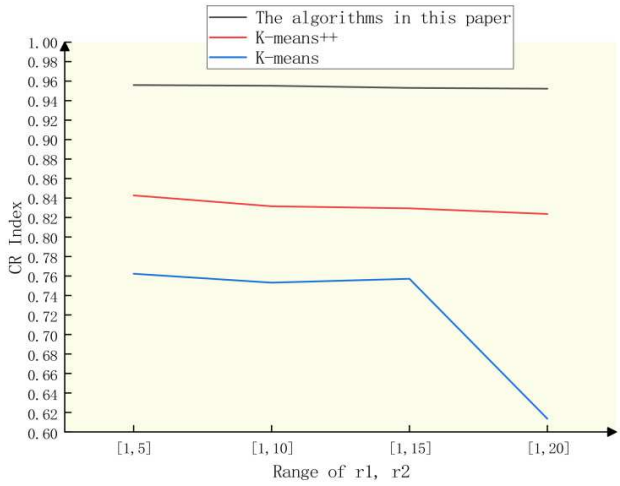

Fig. 5. CR Index Table for Target Clustering of Different Algorithms

The CR index is an indicator used to evaluate the performance of clustering algorithms, commonly known as the Clustering Results Consistency Index. It is used to measure the degree of consistency between different running results when clustering algorithms cluster datasets. The closer the value is to 1, the better the performance; the closer to 0, the more inconsistent the results of the clustering algorithm are. The CR values of this algorithm are listed in Fig. 5. From the CR values shown in Fig. 5, it can be seen that compared to K-means and K-means++algorithms, interval value data dynamic clustering algorithms have better clustering performance and can correctly cluster data into clustering intervals. The correctness and effectiveness of this method have been demonstrated through the analysis of several inter cluster intervals.

TABLE II. COMPARISON OF TARGET GROUPING EFFECTIVENESS

Indicators	The algorithms in this paper	K-means++	K-means
Silhouette score	0.9788	0.8539	0.4885
Accuracy rate/%	100	92.6	74.2

Silhouette score evaluates the performance of the model using the distance between points in the same cluster, as well as the distance between points in the next neighboring cluster and all other points. It is mainly calculated based on factors such as distance, similarity, and compactness between the sample point and its cluster and the nearest neighboring cluster. Silhouette score is an evaluation metric for clustering algorithms, used to measure the quality of clustering results. The value range of this indicator is between 1 and 1, and the closer the value is to 1, the better the clustering result; the closer the value to -1, the worse the clustering result. From Table II, it can be seen that among the Silhouette score index evaluation indicators, the value of the algorithm in this paper was 0.9788, which was the highest. The clustering method based on the K-means algorithm had the lowest score, which was 0.4885, indicating that the clustering method for the data in this paper had the best results. From the analysis of accuracy evaluation indicators, the clustering method based on the algorithm in this paper and the K-mean++algorithm had a relatively high accuracy, with an accuracy of up to 100% based on the algorithm in this paper.

V. CONCLUSION

This article discussed a customer market analysis method based on interval value data dynamic clustering algorithm, and initialized the class of dynamic clustering to quickly initialize the cluster center. At the same time, separating the data improved the parallelism of initialization. This article provided a clustering based data extraction method by analyzing traditional data extraction methods. In the process of type extraction from customer data, using this method can achieve clustering analysis of customers, enabling enterprises to make corresponding marketing decisions, that is, adopting the same marketing strategy for the same type of customers, and adopting differentiated marketing strategies for different customers. This provides a very important theoretical basis for enterprise decision-making, and the application of this concept in practice can also help senior managers of enterprises to better manage the enterprise.

REFERENCES

[1] Abdulhafedh A. Incorporating k-means, hierarchical clustering and pca in customer segmentation[J]. Journal of City and Development, 2021, 3(1): 12-30.

[2] Bandyopadhyay S, Thakur S S, Mandal J K. Product recommendation for e-commerce business by applying principal component analysis (PCA) and K-means clustering: benefit for the society[J]. Innovations in Systems and Software Engineering, 2021, 17(1): 45-52.

[3] Anitha P, Patil M M. RFM model for customer purchase behavior using K-Means algorithm[J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(5): 1785-1792.

[4] Chaudhary K, Alam M, Al-Rakhami M S, et al. Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics[J]. Journal of Big Data, 2021, 8(1): 1-20.

[5] Hopf K, Weigert A, Staake T. Value creation from analytics with limited data: a case study on the retailing of durable consumer goods[J]. Journal of Decision Systems, 2023, 32(2): 289-325.

[6] Sheng J, Amankwah - Amoah J, Khan Z, et al. COVID - 19 pandemic in the new era of big data analytics: Methodological innovations and future research directions[J]. British Journal of Management, 2021, 32(4): 1164-1183.

[7] Ghazal T M. Performances of K-means clustering algorithm with different distance metrics[J]. Intelligent Automation & Soft Computing, 2021, 30(2): 735-742.

[8] Sarbaini S, Saputri W, Muttakin F. Cluster Analysis Menggunakan Algoritma Fuzzy K-Means Untuk Tingkat Pengangguran Di Provinsi Riau[J]. Jurnal Teknologi Dan Manajemen Industri Terapan, 2022, 1(2): 78-84.

[9] Abdullah D, Susilo S, Ahmar A S. The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data[J]. Quality & Quantity, 2022, 56(3): 1283-1291.

[10] Parameshachari, B. D., KM Sunjiv Soyjaudah, and Sumitra Devi KA. "Secure transmission of an image using partial encryption based algorithm." International Journal of Computer Applications 63, no. 16 (2013).

[11] Sinurat M, Heikal M, Simanjuntak A.Product Quality On Consumer Purchase Interest With Customer Satisfaction As A Variable Intervening In Black Online Store High Click Market: Case Study on Customers of the Tebing Tinggi Black Market Online Store[J]. Morfai Journal, 2021, 1(1): 13-21.

[12] Hollebeek L D, Sharma T G, Pandey R. Fifteen years of customer engagement research: a bibliometric and network analysis[J]. Journal of Product & Brand Management, 2022, 31(2): 293-309.

[13] Ali H, Zainal V R, Ilhamalimy R R. Determination of Purchase Decisions and Customer Satisfaction: Analysis of Brand Image and Service Quality (Review Literature of Marketing Management)[J]. Dinasti International Journal of Digital Business Management, 2021, 3(1): 141-153.

[14] Barari M, Ross M, Thaichon S.A meta - analysis of customer engagement behaviour[J]. International Journal of Consumer Studies, 2021, 45(4): 457-477.

[15] Dalmaijer E S, Nord C L, Astle D E. Statistical power for cluster analysis[J]. BMC bioinformatics, 2022, 23(1): 1-28.

[16] Singh S, Kumar K. A study of lean construction and visual management tools through cluster analysis[J]. Ain Shams Engineering Journal, 2021, 12(1): 1153-1162.

[17] Hu L, Zhang J, Pan X. HiSCF: leveraging higher-order structures for clustering analysis in biological networks[J]. Bioinformatics, 2021, 37(4): 542-550.

[18] Li G, Kou G, Peng Y. Heterogeneous large-scale group decision making using fuzzy cluster analysis and its application to emergency response plan selection[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 52(6): 3391-3403.

[19] Karim M R, Beyan O, Zappa A.Deep learning-based clustering approaches for bioinformatics[J]. Briefings in bioinformatics, 2021, 22(1): 393-415.

[20] Parameshachari, B. D. "Big data analytics on weather data: predictive analysis using multi node cluster architecture." International Journal of Computer Applications (2022): 0975-8887.

[21] Gunn R L, Steingrimsson J A, Merrill J E.Characterising patterns of alcohol use among heavy drinkers: A cluster analysis utilising alcohol biosensor data[J]. Drug and Alcohol Review, 2021, 40(7): 1155-1164.

[22] Li X, Guo Y, Zhao T.Cluster analysis of self - concept and job satisfaction in Chinese nurses with master’ s degree to identify their turnover intention: A cross - sectional study[J]. Journal of Clinical Nursing, 2021, 30(13-14): 2057-2067.

[23] Hou J, Sánchez A G, Ross A J. The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: BAO and RSD measurements from anisotropic clustering analysis of the quasar sample in configuration space between redshift 0.8 and 2.2[J]. Monthly Notices of the Royal Astronomical Society, 2021, 500(1): 1201-1221.