The Application of K-MEANS Algorithm-Based Data Mining in Optimizing Marketing Strategies of Tobacco Companies

Mingqian Ma
School of Economics and Management, Jiangsu Vocational College of Finance Economics, Huaian, 223003, China

***Abstract*—With the continuous development of data mining technology, more and more industries are applying data mining techniques to optimize their marketing strategies. In response to the persistent decline in tobacco sales and the gradual erosion of customer base in a particular enterprise in recent years, this study employs data mining technology to enhance the current tobacco marketing strategy. Firstly, in response to the current shortcomings of the company, a marketing optimization design scheme was proposed and a customer classification evaluation index system was constructed. Subsequently, homomorphic encryption technology and enhanced peak density thinking were employed to enhance the conventional K-means algorithm. The enhanced algorithm was then utilized in the customer clustering and partitioning scheme, with the objective of investigating the underlying information present in customer consumption data. The performance of the algorithm was tested, and the results showed that the mean square error of the improved K-means algorithm was about 0.1, with an average absolute error of about 0.05. The highest detection rate in the validation set was 0.95, and the lowest false alarm rate was 0.07. Both experts and customers were highly satisfied with the marketing strategy under the enhanced K-means algorithm. In summary, the clustering analysis method used in this study can effectively uncover the hidden value behind various types of customer data, thereby helping companies to make better marketing strategies.**

***Keywords*—Data mining; homomorphic encryption; k-means; tobacco; marketing strategy; indicator system**

I. Introduction

In today's fiercely competitive market environment, tobacco companies are facing unprecedented challenges and opportunities. In light of evolving consumer preferences and heightened awareness of health concerns, traditional marketing strategies are proving inadequate in meeting market demand. In particular, tobacco companies must contend with not only stringent regulatory constraints but also the necessity of devising bespoke marketing strategies for disparate consumer segments. As a result, accurate customer segmentation and efficient marketing strategies are key to improving sales performance. Currently, many tobacco companies have recognized the importance of data mining (DM) and customer segmentation [1]. By analyzing customer buying patterns and preferences, companies can better understand market demand and develop more accurate marketing plans. However, traditional customer classification methods are often limited by privacy concerns, especially when sensitive information is involved. It has become an urgent problem to solve how to effectively analyze while protecting customer privacy. In addition, although clustering analysis and other techniques have been applied in customer segmentation, existing algorithms still have certain limitations in handling high-dimensional data and ensuring result stability. With the continuous development of DM technology, DM technology using K-means algorithm (KMA) has been extensively applied to the optimization of tobacco companies' marketing strategies [2-3]. In the optimization of marketing strategies for tobacco companies, KMA can help tobacco companies to analyze data such as consumers' consumption habits, taste preferences, and purchasing behavior to develop more scientific and reasonable marketing strategies [4]. In response to this background, this study will use KMA to cluster customer data and identify relevant factors that affect tobacco sales quotas for optimization.

II. Related Work

The KMA is a clustering algorithm in view of distance calculation, which can classify data points with similar features into different clustering centers, thus achieving data analysis and mining. Currently, many scholars have conducted a series of studies on this algorithm [5-6]. Huang B et al. combined the KMA in clustering analysis with moving windows to propose a more effective method for static friction testing. The research results indicated that the method improves the detection accuracy by 15.3% compared to existing static friction detection methods. Applying this technology to practical industrial production could provide effective estimates of static resistance bands, as well as detect severe valve static resistance and unexpected valve closure situations [7]. In response to the severe faults that solar energy may face during use, Et Taleby et al. designed a solar panel detection system using wireless communication technology combined with KMA. The system diagnosed the failure of photovoltaic panels by detecting their thermal images. The outcomes indicated that the detection accuracy of the method used in solar energy fault detection was as high as 96.54%, which had good performance and could diagnose faults in a timely manner to effectively avoid accidents such as fires [8]. Zhang et al. proposed a new encrypted K-means clustering analysis algorithm in a collaborative manner to address the issue of KMA being unable to ensure data privacy during data clustering. In this algorithm, Zhang et al. used secure multi-party computation and differential privacy technology to train the K-means clustering model, aiming to protect data privacy and ensure that data can be output normally for analysis. The outcomes indicated that the adopted method possessed excellent clustering performance,

not only ensuring that data was not leaked during processing, but also updating the cluster center faster to ensure better clustering results [9]. In response to the issue of unequal resource allocation in the current multi-level sales operation network, Sin et al. rebuilt a resource allocation model and achieved the goal of multi-level task automatic allocation through this model. The experiment showed that the sales resource allocation model constructed could automatically allocate personnel and various resources, thereby achieving maximum resource utilization [10]. Currently the construction industry is facing problems such as waste management. It has been the concern of many scholars to find out how to effectively manage waste while maximizing economic benefits. Queheille E et al. proposed a multi-objective optimization algorithm and used the algorithm in waste management in the construction industry, aiming to design a suitable solution through the algorithm, which can maximize the use of human and material resources. The results of the study showed that the multi-objective optimization algorithm used was able to effectively calculate the number of workers required, the type of waste disposal and the optimal management plan, which had a certain value of utilization [11]. In light of the evolving landscape of social media and consumer behavior, the development of effective marketing strategies has emerged as a pivotal concern for major merchants. Gao M et al. used methods such as literature analysis and survey analysis to investigate the sales effect under different operating modes. In addition, multi-objective optimization and backward induction were used to optimize the current product pricing strategy. Finally, the results of the effects of different sales strategies under different network structures were verified through numerical simulation experiments. The experimental results showed that enterprises formulate appropriate marketing strategies could greatly promote the sales of commodities, so as to achieve the purpose of maximizing profitability [12].

In summary, current research on KMA has spread across various fields. In the past, scholars often used KMA to solve problems such as fault diagnosis, data analysis, and resource allocation. At present, no scholars have applied KMA to optimize the marketing strategies of tobacco companies. In response to the current situation of poor tobacco sales performance and frequent customer feedback problems in a certain tobacco company, this study applies KMA to optimize the marketing strategy of the tobacco company. This is aimed at further mining customer information through the KMA to optimize the company's current marketing situation.

III. Research on Marketing Strategies of Tobacco Companies Based on Homomorphic Encryption KMA

For optimizing the current marketing strategy of a tobacco company and improve its sales performance, this study takes a listed tobacco company as an example, first analyzes its current sales problems, proposes a series of improvement measures, and constructs a customer classification evaluation index system. Then it uses the improved peak density KMA under homomorphic encryption technology to evaluate the customer classification evaluation index system, and designs the final customer clustering division scheme. This is to further analyze customer consumption data to optimize marketing strategies.

A. Optimization Design of Marketing Strategy for Tobacco Companies

Homomorphic encryption is an encryption technique that allows the computation of ciphertext without decrypting the data, and ensures that the computed results are consistent with the original plaintext after decryption. Peak density is a concept commonly used to analyze signal or data set features. It describes the density of peak values in a signal or data set and is typically used to measure the distribution of a feature in the data. Tobacco marketing refers to the use of marketing strategies by tobacco companies to increase product sales and market share. In the tobacco company's marketing strategy, its composition includes the research on the target market, customer needs and competitors, as well as the development of personalized Marketing plan [13]. The goal of tobacco marketing is to increase brand awareness, customer satisfaction, and sales revenue to meet customer needs. This study takes a listed tobacco company as an example. In response to its current poor tobacco sales strategy and low sales revenue, a series of improvement and optimization measures are first proposed. Then, clustering analysis algorithm is used to mine data on its tobacco consumer types, to further improve the current tobacco strategy and increase the company's tobacco sales in view of different customer consumption situations. Fig. 1 shows the marketing strategy optimization path of a tobacco company.

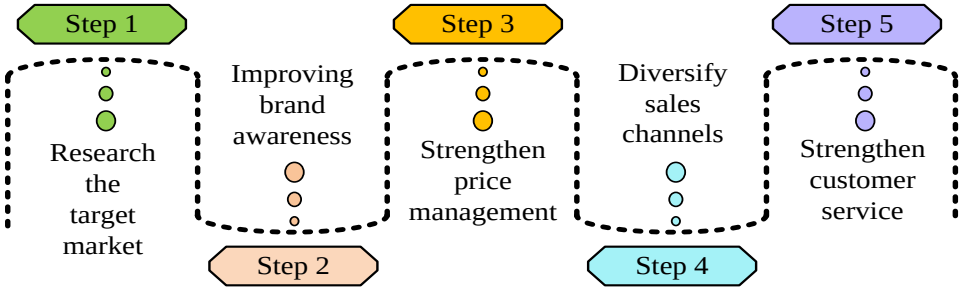


Fig. 1. Optimization path of Tobacco Company's marketing strategy.

Fig. 1 shows the optimization path of the tobacco company's marketing strategy. The entire optimization path consists of five parts, namely, researching the target market, improving brand awareness, strengthening price management, diversifying distribution channels, and strengthening customer service. The purpose of studying the target market is to determine consumer needs so that the company's overall tobacco marketing goals are in line with consumer needs. Improving brand awareness means creating a brand effect. Strengthening price management means that management should formulate reasonable pricing policies to ensure the stability of brand value. With the development of the Internet, traditional offline sales are no longer the only way for people to consume. Diversified sales channels require merchants to combine various online and offline channels for marketing to meet the needs of different consumers. Strengthening customer service requires improving customer satisfaction with their purchases, thereby enhancing their purchasing opinions. Among the above optimization measures, the long-term stable tobacco purchasing behavior of customers is an important guarantee for tobacco sales profitability.

For better understanding customer needs and develop appropriate tobacco sales plans, it is necessary to use DM technology to further mine the current customer information, discover the hidden value behind the data, and effectively improve the current sales model. Customer segmentation theory is a theory that categorizes customers according to certain criteria, to better understand their characteristics, needs, and preferences, and distinguish different categories of customers. Customer segmentation theory can help companies better understand customer needs and behaviors, thereby developing more personalized and effective marketing strategies and improving customer satisfaction and loyalty. In customer segmentation theory, key indicators such as customer purchase frequency, purchase amount, preference level, etc. are usually used to evaluate customer demand for a company's products and services. This study first utilized customer segmentation theory to construct a customer classification index system. Next, it uses the indicator system as an evaluation standard and uses clustering analysis algorithms to study the relationship between various types of customer behavior characteristics and their underlying purchasing behavior. Table I shows the customer classification evaluation index system.

Table I shows the customer classification evaluation index system. Table I shows that the entire customer classification evaluation index system consists of three primary indicators and 13 secondary indicators. According to the customer segmentation theory, customers are classified into three primary indicators: customer value, customer characteristics, and customer behavior. Customer value is further divided into three secondary indicators: total sales quota, total business quota, and average sales price. Customer characteristics are subdivided into five secondary indicators: business location, business market, business form, business nature, and business scale. Customer behavior is divided into five secondary indicators in view of customers' preferences for purchasing tobacco: first class tobacco proportion, second class tobacco proportion, third class tobacco proportion, fourth class tobacco proportion, and fifth class tobacco proportion. The proportion of cigarettes in one category refers to the proportion of customers who purchase tobacco for a single amount between 80 and 100 yuan. The proportion of second-class cigarettes refers to the proportion of customers who buy tobacco for a single amount of 60-80 yuan. The proportion of three kinds of cigarettes refers to the proportion of customers who buy tobacco for a single amount of 40-60 yuan. The proportion of four kinds of cigarettes refers to the proportion of customers who buy tobacco for a single amount of 20-40 yuan. The proportion of five kinds of cigarettes refers to the proportion of customers who buy tobacco for a single amount of less than 20 yuan.

TABLE I. CUSTOMER CLASSIFICATION EVALUATION INDEX SYSTEM

Evaluation Indicator System	First-level indicators	Second-level indicators	Code
Customer classification evaluation index system	Customer Value	Total sales limit	Y1
		Total business limit	Y2
		Average sales price	Y3
	Customer Characteristics	Business location	Y4
		Operating the market	Y5
		Business form	Y6
		Business nature	Y7
		Business scale	Y8
	Customer Behavior	Buy Class I cigarettes	Y9
		Buy Class II cigarettes	Y10
		Buy Class III cigarettes	Y11
		Buy Class IV cigarettes	Y12
		Buy Class V cigarettes	Y13

B. Construction of Tobacco Consumer Segmentation Model Based on Homomorphic Encryption KMA

Homomorphic encryption technology is a cryptographic technology, which enables encrypted objects to be calculated without revealing any useful information [14-15]. In addition, homomorphic encryption can convert encrypted data into the same form as ordinary data, and then perform calculations, so that users can get the expected calculation results even if they do not know the contents of encrypted data. This research applies the homomorphic encryption technology to the KMA, aiming to ensure that the algorithm can obtain information encryption in the running process, so as to prevent data information leakage. The homomorphic encryption process is shown in Fig. 2.

Fig. 2 shows the homomorphic encryption process. Firstly, it is necessary to generate a key suitable for encryption. Keys can be generated using non-public algorithms, such as secret sharing or common algorithms used to solve mathematical Hard problem of consciousness. Next, it uses the key to encrypt the data. This step can be done using a technology called homomorphic encryption, which converts encrypted data into the same form as ordinary data and then encrypts it. In the case of homomorphic encryption, anyone can compute the encrypted data without knowing the key. The ciphertext operation can be achieved by using machine learning algorithms.. Finally, the user needs to receive the decrypted plaintext and provide feedback on the encryption result to determine whether the desired result has been obtained. The encryption process of ciphertext is shown in Eq. (1) to Eq. (3) [16].

$$C = \left(c_1, c_2, \cdots c_n\right) \tag{1}$$

In Eq. (1), C represents the ciphertext data set. c_1, c_2 and c_n represent the ciphertext in the ciphertext dataset, respectively [17-18].

$$M = \left(m_1, m_2, \cdots m_n\right) \tag{2}$$

In Eq. (2), M represents the set of plaintext data. m_1, m_2 and m_n represent the plaintext in the plaintext dataset, respectively. The plaintext encryption process obtained by combining Eq. (1) and Eq. (2) is shown in Eq. (3).

$$\begin{cases} c_1 = Enc\left(m_1\right) \\ c_2 = Enc\left(m_2\right) \end{cases} \tag{3}$$

In Eq. (3), $Enc(\cdot)$ represents the encryption algorithm. Clear text m_1 and m_2 can obtain corresponding ciphertext through encryption algorithms, denoted as c_1 and c_2.

$$c_1 \odot c_2 = Enc(m_1) \odot Enc(m_2) = Enc(m_1 \odot m_2) \tag{4}$$

In Eq. (4), $\odot$represents an effective algorithm. If Eq. (4) satisfies the pre - and post equality relationship under the operation of an effective algorithm, then $Enc(\cdot)$ is said to have homomorphism.

$$Enc\left(m_1\right) \oplus Enc\left(m_2\right) = Enc\left(m_1 + m_2\right) \tag{5}$$

Eq. (5) is the calculation formula of addition homomorphic encryption. $\oplus$ represents an addition operation, and if Eq. (5) is satisfied, it indicates that the encryption algorithm used has additive homomorphism.

$$Enc\left(m_1\right) \otimes Enc\left(m_2\right) = Enc\left(m_1 \otimes m_2\right) \tag{6}$$

Eq. (6) is the calculation formula of multiplicative homomorphic encryption. $\otimes$ represents multiplication operation. Moreover, if Eq. (6) is satisfied, it indicates that the encryption algorithm used has multiplicative homomorphism [19-20].

$$Enc\left(m_1\right) \otimes m_2 = Enc\left(m_1 \times m_2\right) \tag{7}$$

Eq. (7) is the calculation formula of mixed multiplication homomorphic encryption. If equation (7) is satisfied, it indicates that the encryption algorithm used has mixed multiplicative homomorphism.

The KMA is a common clustering algorithm that can assign data objects to the clusters they belong to, so that each object has the same category throughout the entire dataset [21]. The traditional KMA is generally divided into several steps: initializing parameters, clustering, adjusting clustering, and repeating clustering. The clustering process is shown in Fig. 3.

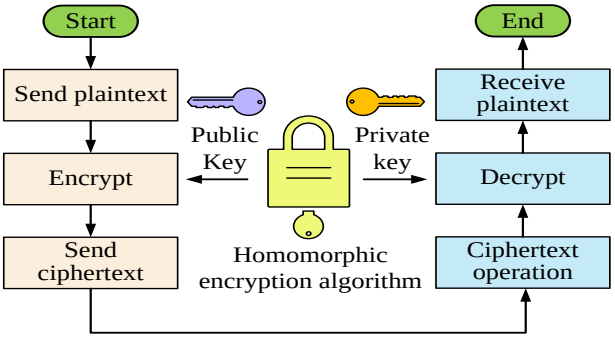


Fig. 2. Homomorphic encryption process.

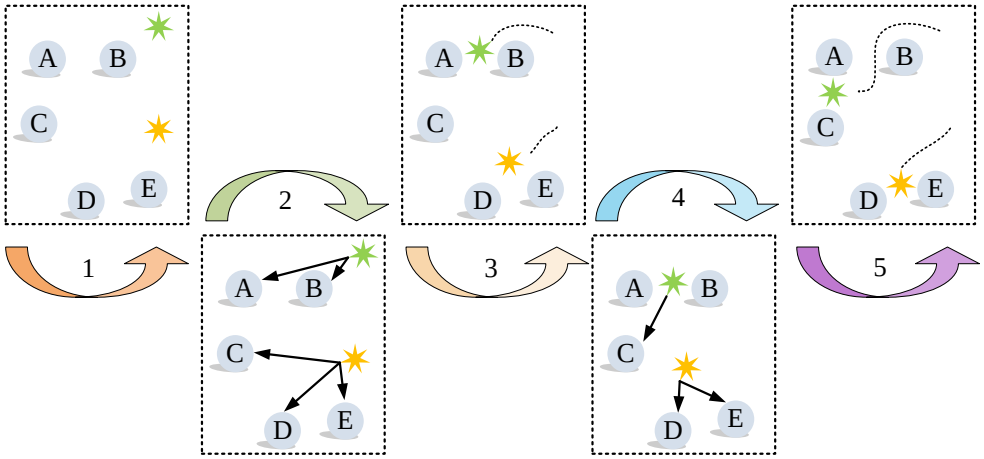


Fig. 3. KMA clustering process.

Fig. 3 shows the clustering process of the KMA. Fig. 3 shows that in the K-means clustering, the initial clustering needs to be selected first. The KMA uses a random selection method to determine the position of each data object in the dataset, and then clusters the data objects and assigns them to the corresponding clusters [22-23]. Over time, the size of the cluster is continuously adjusted to ensure that each object has the same category throughout the entire dataset. It continuously repeats the KMA for generating the best clustering [24-25]. The KMA generally utilizes Euclidean distance for measuring the similarity between two samples, and its mathematical expression is shown in Eq. (8).

$$D(X,Y)=\sqrt{\sum_{t=1}^{m}(x_i-y_i)^2} \tag{8}$$

In Eq. (8), assuming that the dataset has n samples of the m dimension, the distance expression between any two samples X and Y is denoted as $D(X,Y)$. Among them, $i \in 1,2,\mathrm{L},m$.

$$u_i=\frac{1}{|C_i|}\sum_{j=1}^{|C_i|}X_j=\frac{1}{|C_i|}\left(X_1+X_2+\mathrm{L}+X_{|C_i|}\right) \tag{9}$$

Eq. (9) is the iterative calculation formula for the centroid u_i in the KMA. C_i represents the i-th partition cluster. $|C_i|$ represents the quantity of samples contained in the i-th partition cluster.

$$SSE=\sum_{i=1}^{k}\sum_{j=1}^{|C_i|}\left(d\left(X_{i,j},u_i\right)\right)^2 \tag{10}$$

Eq. (10) is the mathematical expression for the standard function SSE of the sum of squared errors. The value of the error squared sum standard function can determine whether the algorithm iteration has ended. When the sum of squares of errors is less than the set error, the algorithm is terminated. k represents dividing the sample into Class k. $X_{i,j}$ represents the j-th sample in the i-th cluster. $d\left(X_{i,j},u_i\right)$ represents the distance from the sample to the center of mass.

$$CPD=\frac{1}{k}\sum_{i=1}^{k}d\left(u_i,u_i'\right) \tag{11}$$

Eq. (11) is the formula for calculating the difference in centroid changes. In addition to using SSE for judgment, this formula can also be used to determine whether the algorithm has ended. u_i' represents the position of the centroid of the previous generation in the i-th cluster. u_i represents the position of the current centroid. $d\left(u_i,u_i'\right)$ represents the Euclidean distance between two centroids. If the difference in centroid changes meets the requirements, the algorithm will be terminated.

Although the K-means clustering analysis algorithm has the characteristics of simple operation, high stability, and good analysis results, it still has certain limitations when used in different scenarios. For example, it has a strong dependence on the k-value of the divided sample category, and the algorithm is greatly affected by the center point. The algorithm has high randomness, making it difficult to converge to the optimal solution state. In response to the above issues, this study attempts to enhance the sample density in the algorithm dataset and uses the principle of maximum density to select the center point. Finally, it proposes a KMA with improved peak density to improve the stability of the traditional KMA and accelerate its Rate of convergence.

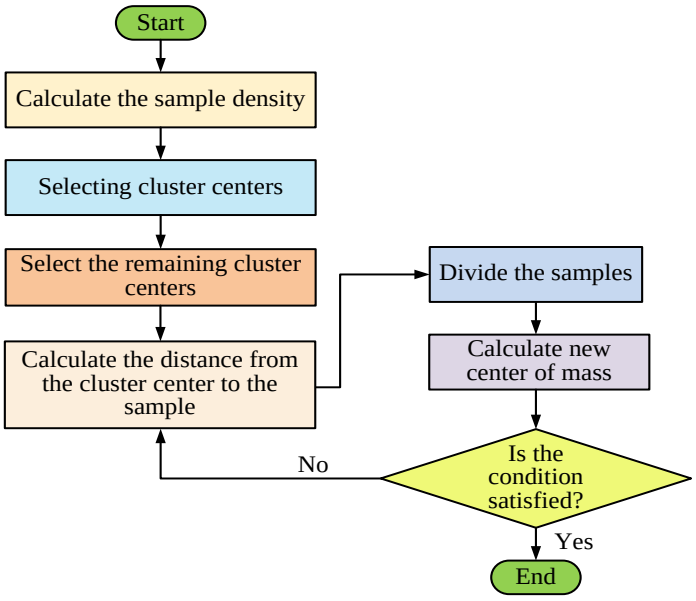


Fig. 4. Flow chart of KMA for improving peak density.

Fig. 4 shows the flowchart of the KMA for improving peak density. The input of the whole algorithm includes the original dataset D, the half price parameter q, the number of clusters k and the error parameter e. The output result is the divided clustering result. Firstly, it sets r and calculates the sample density, recording the sample with the highest density as x_i and the corresponding density value of the sample as $p(x_i)$. It places x_i into set x and removes other sample values within its radius range from the dataset, performing this operation on all samples in the dataset until the initial cluster center set x is obtained. It selects the sample with the highest density as the first clustering center, and then selects the second sample with the farthest distance as the second clustering center, using this method to select the remaining initial centers. Next, it uses the distance formula mentioned above to calculate the distance between the sample and the center, and divides the samples according to the principle of minimum distance. Then it calculates the new centroid for each classification cluster in the partitioned dataset using Eq. (9). Finally, it determines whether the output result is less than the error parameter e in view of the convergence judgment formula. If it is true, the algorithm is terminated. The improved peak density KMA is clustered using homomorphic encryption technology, and the customer segmentation scheme is shown in Fig. 5.

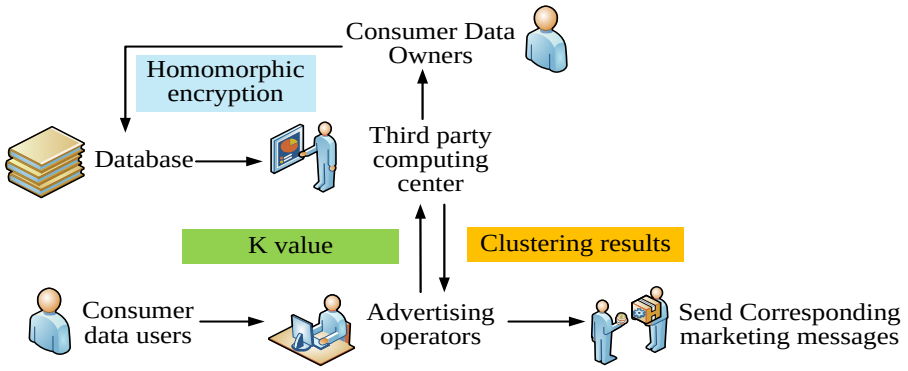


Fig. 5. Design diagram of customer clustering and division scheme.

Fig. 5 shows the design diagram of the customer clustering division scheme. To ensure the maximum utilization of consumption data, it is first necessary for data users to provide consumption data of their stores to advertising operators. However, to ensure the privacy of data information, the system will introduce homomorphic encryption technology to encrypt information during the entire data transmission process. Then it shares the data information with the third-party computing center, so as to use the KMA to analyze the customer consumption data, and then generate the corresponding marketing messages.

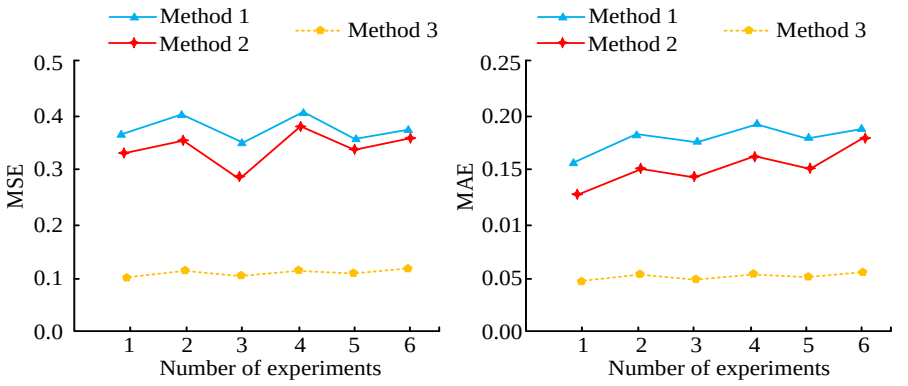
IV. Analysis of Tobacco Company Marketing Strategy Optimization Effect Based on Homomorphic Encryption KMA

For testing the effectiveness of the methods used in the research, the results analysis part first tested the performance of the improved peak density KMA under homomorphic encryption. Moreover, it proves that the improved peak density

KMA performs better than other comparative algorithms in four indicators: detection rate, false alarm rate, accuracy, and clustering running time. Subsequently, this study applied the improved peak density KMA to analyze customer consumption data, validated the customer classification evaluation index system, and obtained the application effect of tobacco company marketing strategy optimization.

A. *Performance Test of Improved Peak Density KMA Under Homomorphic Encryption*

It selected the tobacco sales data of a listed tobacco company in the past three years as the dataset for this experiment. After simple preprocessing of the dataset, the remaining data is divided into a training set and a validation set in a 9:1 ratio to test the performance of different clustering algorithms. The mean squared error (MSE) and mean absolute error (MAE) of the traditional KMA (subsequently recorded as method 1), the Mean shift algorithm (subsequently recorded as method 2), and the improved peak density KMA (subsequently recorded as method 3) in the same tobacco dataset are shown in Fig. 6.

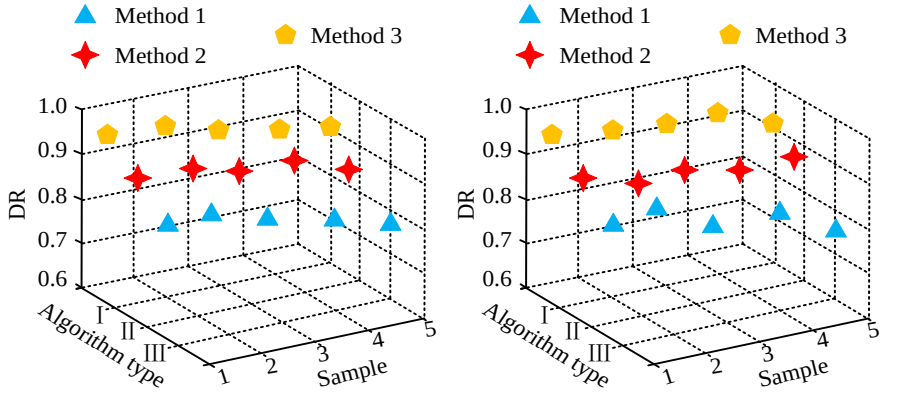


(a) MSE of each algorithm model under different experimental times

(b) MAE of each algorithm model under different experimental times

Fig. 6. MSE and MAE changes of three clustering methods.

Fig. 6 (a) shows the MSE changes of the three methods under different experimental times. As the number of experiments increases, the MSE of Method 1 and Method 2 fluctuates up and down, with a large fluctuation range, while the MSE of Method 3 is stable at around 0.1. Fig. 6 (b) shows the MAE changes of the three methods under different experimental times. The MAE values of Method 1 and Method 2 have significant changes, while the MAE values of Method 3 are stable at around 0.05.

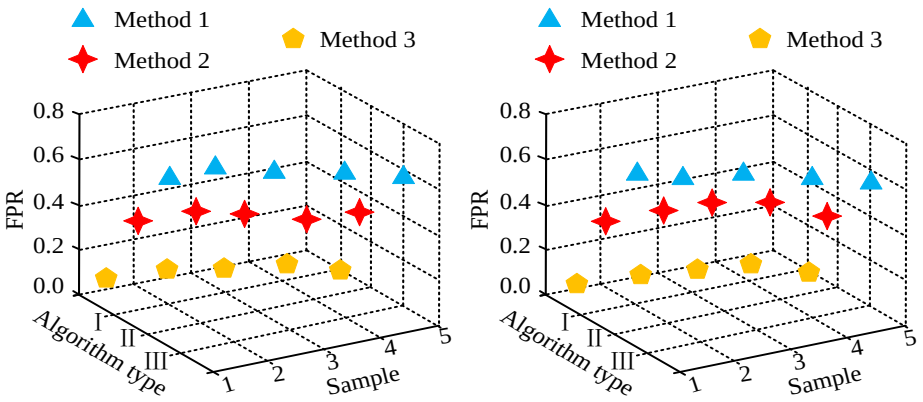


(a) Detection rate of different types of algorithms in training sets

(b) Detection rate of different types of algorithms in the test set

Fig. 7. DR changes of three clustering methods.

Fig. 7 (a) and 7 (b) show the changes in detection rates of the three clustering methods in the training and validation sets, respectively. Fig. 7 (a) shows that the highest detection rates of Method 1, Method 2, and Method 3 in the training set are 0.76, 0.85, and 0.97, respectively. Fig. 7 (b) shows that the highest detection rates of Method 1, Method 2, and Method 3 in the validation set are 0.77, 0.88, and 0.95, respectively. In summary, method three has the best data detection effect.

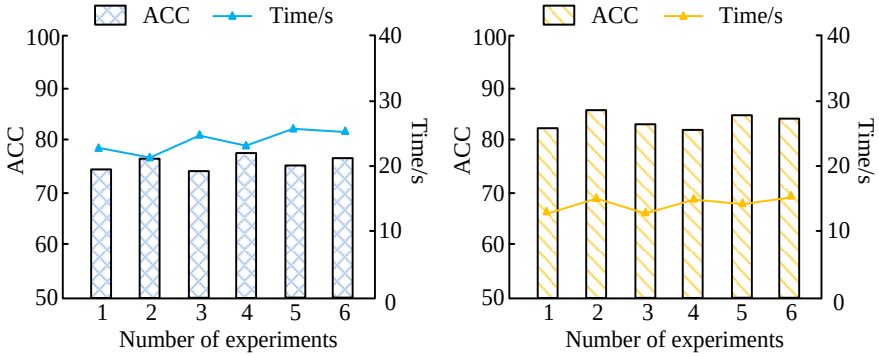


(a) False Positive Rate of different types of algorithms in training sets

(b) False Positive Rate of different types of algorithms in the test set

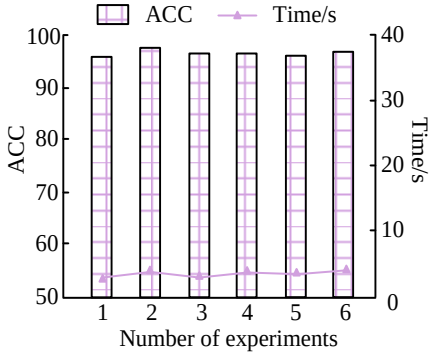
Fig. 8. FPR variation of three clustering methods.

Fig. 8 (a) and Fig. 8 (b) show the changes in false positive rates of the three clustering methods in the training and validation sets, respectively. Fig. 8 (a) shows that the highest false positive rates of Method 1, Method 2, and Method 3 in the training set are 0.56, 0.38, and 0.09, respectively. Fig. 8 (b) shows that the highest false alarm rates for Method 1, Method 2, and Method 3 in the validation set are 0.52, 0.39, and 0.07, respectively. In summary, method three has the best false alarm rate performance.



(a) The clustering accuracy and runtime of the method 1

(b) The clustering accuracy and runtime of the method 2



(c) The clustering accuracy and runtime of the method 3

Fig. 9. ACC and time variation of three clustering methods.

Fig. 9 (a), 9 (b), and 9 (c) respectively show the clustering accuracy and runtime changes of the three clustering methods under multiple experiments. Figure 9 (a) shows that the average clustering accuracy of Method 1 in six experiments is 73.26. As the number of experiments changes, its running time always fluctuates between 20 and 30 seconds, with significant fluctuations. Fig. 9 (b) shows that the average clustering accuracy of Method 2 in six experiments is 84.12. As the number of experiments increases, its running time always fluctuates between 10-20 seconds, with a smaller fluctuation compared to Method 1. Fig. 9 (c) shows that the average clustering accuracy of Method 3 in six experiments is 96.08, and its running time remains around 4 seconds as the number of experiments increases.

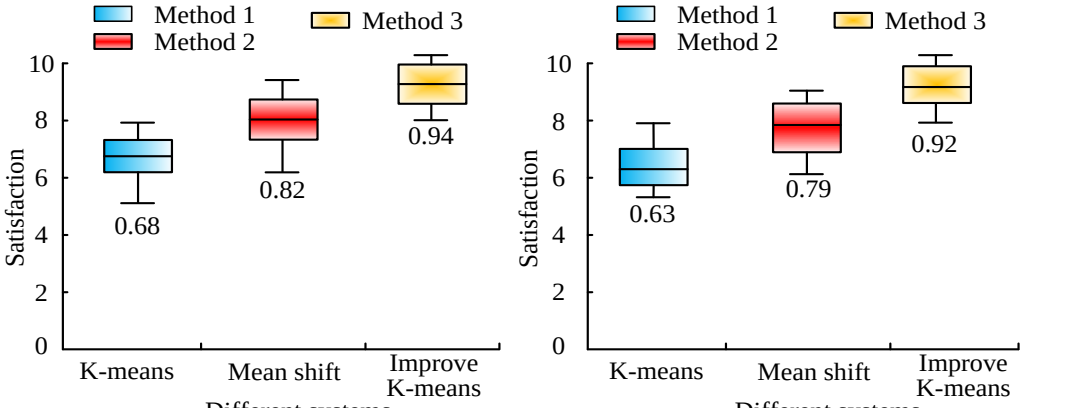
B. Analysis of the Optimization Effect of Tobacco Company's Marketing Strategy

After testing the performance of the improved peak density KMA under homomorphic encryption, the research applied it to the customer classification evaluation index system, input various customer data into the model, and obtained the final output results as shown in Table II. By examining the results of various outputs, it is possible to assist staff in the implementation of corresponding marketing plans, thereby achieving the goal of optimizing marketing plans and increasing sales.

TABLE II. OUTPUT RESULTS OF CUSTOMER CLASSIFICATION EVALUATION INDICATORS

Evaluation indicator system	First-level indicators	Second-level indicators	Code	Output value
Customer classification evaluation index system	Customer value	Total sales limit	Y1	8.7
		Total business limit	Y2	8.8
		Average sales price	Y3	9.3
	Customer Characteristics	Business location	Y4	9.2
		Operating the market	Y5	8.5
		Business form	Y6	8.6
		Business nature	Y7	9.0
		Business scale	Y8	8.3
	Customer behavior	Buy Class I cigarettes	Y9	7.7
		Buy Class II cigarettes	Y10	8.1
		Buy Class III cigarettes	Y11	8.8
		Buy Class IV cigarettes	Y12	8.7
		Buy Class V cigarettes	Y13	7.6

Table II shows the output results of customer classification evaluation indicators. Table II shows that out of the 13 secondary indicators, a total of 3 secondary indicators have output values above 9 points, namely average sales price, business location, and business nature. In addition, 8 secondary indicators have an initial value of more than 8 points. Given the above scoring results, tobacco company executives should optimize the drivers with higher scores to increase the company's tobacco revenues.



（a）User satisfaction with marketing strategies for various systems （b）Expert satisfaction with marketing strategies for various systems

Fig. 10. Satisfaction of users and experts with different marketing schemes.

Fig. 10 shows the satisfaction of users and experts with different marketing plans. Fig. 10 (a) shows the satisfaction of users with marketing solutions under different clustering methods. Fig. 10 (a) shows that users' satisfaction with marketing solutions under the three clustering methods of K-means, Mean shift, and Improve K-means is 0.68, 0.82, and 0.94, respectively. Fig. 10 (b) shows the satisfaction of experts with marketing solutions under different clustering methods. Fig. 10 (b) shows that experts' satisfaction with marketing solutions under the three clustering methods of K-means, Mean shift, and Improve K-means is 0.63, 0.79, and 0.92, respectively.

V. DISCUSSION

The research focused on optimizing the marketing strategies of tobacco companies, and achieved efficient customer classification and privacy protection through the improved peak

density KMA with homomorphic encryption. In practice, the marketing challenges faced by tobacco companies were mainly due to the intensification of market competition and the diversification of consumer demand. Traditional marketing methods were difficult to accurately reach target customers, resulting in persistently low sales. Therefore, using modern DM techniques for customer segmentation could not only help companies identify the characteristics of different customer groups, but also develop more personalized marketing strategies based on these characteristics. The experimental results showed that the improved KMA outperformed the traditional methods in clustering performance. In the experiment, the detection rate of the improved algorithm reached 97%, while the detection rates of the traditional KMA and the mean shift algorithm were 76% and 85%, respectively. This gap indicated the effectiveness of the improved algorithms in handling complex data sets, especially in capturing subtle differences in customer behavior. In addition, the false positive rate of the enhanced algorithm was only 9%, which was significantly lower than the 56% and 38% of the traditional algorithms. This improved accuracy was due to the optimization of the algorithm in the initial cluster center selection, which reduced the problem of inaccurate clustering caused by improper center selection. The introduction of homomorphic encryption technology effectively solved the privacy problem and ensured the security of customer information during the analysis process, which was particularly important for the tobacco industry. Due to increasingly stringent industry regulations, protecting customer privacy become a key factor in the sustainable development of companies. For example, data breaches could result in significant fines and loss of brand reputation. Therefore, the adoption of homomorphic encryption technology to ensure the privacy of customer data during the analysis process was a necessary choice in line with industry trends. In conclusion, the improved peak density KMA based on homomorphic encryption provided an effective solution for optimizing the marketing strategies of tobacco companies. On this basis, future research could further explore other data processing techniques and algorithms to cope with increasingly complex market environments and customer demands.

VI. Conclusion

In response to the problems of a certain tobacco company's current sales quota not being ideal and a large number of users losing, this study used the KMA in cluster analysis to analyze its customer data and optimize its current marketing strategy. The research results indicated that the improved KMA used had good clustering performance, with MSE and MAE values fluctuating around 0.1 and 0.05, respectively, which was much smaller than the traditional KMA and Mean shift algorithm. The maximum detection rate of improved K-means in the validation set was 0.95, the minimum false alarm rate was 0.07, and the average clustering accuracy under six experiments was 96.08. The clustering time was maintained at around four seconds, and its performance was superior to the other two comparative algorithms. In addition, the improved KMA had high output values for average sales price, business location, and business nature. Therefore, targeted optimization was needed for these three secondary indicators. In the end, both experts and users had a satisfaction level of over 0.9 with the optimized marketing plan. In summary, the improved K-means method adopted in this study could better develop appropriate marketing plans in view of consumer data characteristics, thereby obtaining the satisfaction of users and experts, and helping the company achieve maximum profitability. However, due to only analyzing customer impact indicators, there was still some error. Future research directions could explore more influencing factors, not limited to the customer. In addition to customer impact indicators, future research may wish to consider the inclusion of multidimensional variables such as market environment, competitor behavior, and economic factors in order to provide a more comprehensive evaluation of their impact on sales. Furthermore, a hybrid approach combining deep learning techniques with traditional clustering algorithms may be employed to enhance the accuracy and adaptability of the model.

References

[1] Jeong W, Almubarak M S, Tsingas C. Quality control for the geophone reorientation of ocean bottom seismic data using k-means clustering. Geophysical Prospecting, 2021, 69(7):1487-1502.

[2] Fan Y, Liu Y, Qi H, Liu F, Ji X J. Anti-Interference Technology of Surface Acoustic Wave Sensor Based on K-Means Clustering Algorithm. IEEE Sensors Journal, 2021, 21(7):8998-9007.

[3] Pan S, Yan K, Yang H, Jiang C, Qin Z. A sparse spike deconvolution method based on Recurrent Neural Network like improved Iterative Shrinkage Thresholding Algorithm. Geophysical Prospecting for Petroleum, 2022, 58(4):533-540.

[4] Gao S, Gao S, Pan W, Wang M. Design of Improved PID Controller Based on PSO-GA Hybrid Optimization Algorithm in Vehicle Lateral Control. Studies in informatics and control, 2021, 30(4):55-65.

[5] Liu S, Sun L, Zhu S, Li J, Chen X, Zhong W. Operation strategy optimization of desulfurization system based on data mining. Applied Mathematical Modelling, 2020, 81(5):144-158.

[6] Niu X, Wang J. A combined model based on data preprocessing strategy and multi-objective optimization algorithm for short-term wind speed forecasting. Applied Energy, 2019, 241(5):519-539.

[7] Huang B, Zheng D, Sun X, Amalraj J, Shah A, Damarla S K. Valve Stiction Detection and Quantification Using a K-Means Clustering Based Moving Window Approach. Industrial & Engineering Chemistry Research, 2021, 60(6):2563-2577.

[8] Et-Taleby A, Chaibi Y, Boussetta M, Allouhi A, Benslimane M. A novel fault detection technique for PV systems based on the K-means algorithm, coded wireless Orthogonal Frequency Division Multiplexing and thermal image processing techniques. Solar Energy, 2022, 237(5):365-376.

[9] Zhang E, Li H, Huang Y, Hong S, Zhao L, Ji C. Practical Multi-party Private Collaborative k-means Clustering. Neurocomputing, 2022, 467(1):256-265.

[10] Sin H H, Parlar M. Modeling and optimization of multilevel marketing operations. Naval Research Logistics (NRL), 2022, 69(4):581-598.

[11] Queheille E, Taillandier F, Saiyouri N. Optimization of strategy planning for building deconstruction. Automation in Construction, 2019, 98(2):236-247.

[12] Gao M, Zhao M, Qin J. Marketing strategy selection of referral reward under social network: RRWS, RRMS, or dual strategy? International Transactions in Operational Research, 2022, 29(3):1970-2001.

[13] Hall J, Cho H D, Guo Y, Mildred M, Thompson L A, Shenkman E, Salloum R.Association of Rates of Smoking During Pregnancy With Corporate Tobacco Sales Policies. Jama Pediatrics, 2019, 173(3):284-286.

[14] Guo Y, Mustafaoglu Z, Koundal D. Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. Journal of Computational and Cognitive Engineering, 2022, 2(1):5-9.

[15] Zhao X, Nie F, Wang R, Li X. Improving projected fuzzy K-means clustering via robust learning. Neurocomputing, 2022, 491(6):34-43.

[16] Bingqian Z, Xingsheng G. Multi-block statistics local kernel principal component analysis algorithm and its application in nonlinear process fault detection. Neurocomputing, 2020, 376(2):222-231.

[17] Ghaffari R, Golpardaz M, Helfroush M S, Danyali H. A fast, weighted CRF algorithm based on a two-step superpixel generation for SAR image segmentation. International Journal of Remote Sensing, 2020, 41(9):3535-3557.

[18] Mano A, Anand S. Method of multi-region tumour segmentation in brain MRI images using grid-based segmentation and weighted bee swarm optimisation. IET Image Processing, 2020, 14(12):2901-2910.

[19] Dickinson T A, Richman M B, Furtado J C. Subseasonal to Seasonal Extreme Precipitation Events in the Contiguous United States: Generation of a Database and Climatology. Journal of Climate, 2021, 34(18):7571-7586.

[20] Lee E H, Kim K, Kho S Y, Kim D K, Cho S H. Estimating Express Train Preference of Urban Railway Passengers Based on Extreme Gradient Boosting (XGBoost) using Smart Card Data.Transportation Research Record, 2021, 2675(11):64-76.

[21] Tao T, Liu Y, Qiao Y, Gao L , Lu J, Zhang C, Wang Y. Wind turbine blade icing diagnosis using hybrid features and Stacked-XGBoost algorithm. Renewable Energy, 2021, 180(12):1004-1013.

[22] Sagi O, Rokach L. Approximating XGBoost with an interpretable decision tree. Information Sciences, 2021, 572(2):522-542.

[23] Zhao X, Nie F, Wang R, Li X. Improving projected fuzzy K-means clustering via robust learning. Neurocomputing, 2022, 491(6):34-43.

[24] Changhua L, Yanxia Z, Chenzhou C. Identification of BASS DR3 sources as stars, galaxies, and quasars by XGBoost. Monthly Notices of the Royal Astronomical Society, 2021, 506(2):1651-1664.

[25] Raichura M, Chothani N, Patel D. Efficient CNN-XGBoost technique for classification of power transformer internal faults against various abnormal conditions. IET Generation, Transmission & Distribution, 2021, 15(5):972-985.