Enhanced E-Commerce Data Processing Using Dimensional Control and Optimized KNN Algorithm

Ling Yang, Fuli Qi*
School of Information Engineering, Shanghai Zhongqiao Vocational and Technical University, Shanghai, 200000, China
E-mail: qifuli2008@163.com
*Corresponding author

Keywords: data classification, data mining, E-commerce big data, artificial intelligence, KNN algorithm, K value selection policy

Received: May 29, 2024

To solve the current problem of low accuracy and time-consuming data mining and classification techniques applied to e-commerce platforms, the study proposes an e-commerce data processing model based on data mining and improved KNN classification algorithm. The model first uses dimensional control mechanism and Spark mechanism together to deeply mine the massive e-commerce data. Subsequently, it utilizes KNN algorithm based on K-value selection strategy to classify the mined data. The performance comparison experiment of data mining algorithms shows that the mining time of the proposed data mining algorithm is 4.6 min, and the mining error rate was 4.2%. Compared with the other two algorithms, the mining time was reduced by about 50% and the error rate was reduced by about 50%. Comparative experiments on the improved KNN algorithm showed that the classification recognition rate of KNN algorithm based on K-value selection strategy was 97.3%, and the classification time was 27.3 s. Compared with traditional KNN algorithm and KNN algorithm based on K-means clustering, the classification recognition rate was increased by about 3%, and the classification time was shortened by more than 90%. The above results show that the proposed method can not only improve the accuracy of e-commerce data classification, provide data support for the precision marketing of e-commerce platforms, but also provide new ideas for the strategic transformation of e-commerce platforms.

Povzetek: Za obdelavo podatkov e-trgovine je izboljšanim algoritem KNN, ki z mehanizmom dimenzijske kontrole doseže boljše razvrščanje ter zmanjšuje čas obdelave.

1 Introduction

With the rapid development of the data era, the scale of e-commerce platforms is getting bigger and bigger [1]. For e-commerce platforms with an increasing number of users, it is of great significance to better analyze and predict user behavior and to carry out accurate marketing to users on this basis [2]. Data mining algorithms are mainly a trial method and calculation for building data mining models based on the data provided, with the aim of extracting the required data from the dataset [3]. Currently, there are many e-commerce platforms that apply data mining techniques, but the results are not ideal [4]. The KNN algorithm is a common classification algorithm, which is often used to classify products and users in e-commerce platforms. However, its classification effect is not ideal either [5]. To make better use of e-commerce big data and accelerate the strategic transformation of e-commerce platforms, the study uses an optimized KNN classification algorithm to classify the data obtained from data mining algorithms.

2 Review of the literature

The vigorous development of the age of big data has driven the wide application of data mining algorithms and classification algorithms, which have been increasingly used in various fields. For example, in order to conduct data mining on the energy consumption time series data set of primary school buildings, Gong et al. used k-shape algorithm and Apriori algorithm to conduct clustering analysis and association analysis on the energy consumption time series data. The results showed that the curve obtained by the data mining algorithm could effectively translate the operation characteristics of primary school buildings [6]. To better ensure the safe and stable operation of large heating systems, Huang et al. proposed a data mining method combining professional knowledge, through which the on-demand parameters of large heating systems can be accurately predicted. The empirical analysis of this method showed that this method could obtain accurate prediction results from the historical data of heat sources, and could effectively ensure the operation of large heating systems [7]. Zare et al. proposed a model based on a neural network algorithm

to address the challenge of determining the relative importance of the contributing factors to industrial workers' hearing loss. They employed the model for empirical analysis. The results showed that the model could help workers better identify the influencing factors of hearing loss. It provided data support for the implementation of hearing protection plans in the industry [8]. Becker et al. used data mining algorithm to analyze the differences between capillary and venous blood technical analytes in order to explain the capillary test results more reasonably. The results showed that the results obtained through this data mining algorithm improved the guidance for children's capillary test results interpretation [9]. Sha et al. put forward a data mining and fusion model based on Raman spectrum and mid infrared spectrum to solve the problem of insufficient rice geographical origin recognition ability. The model was empirically derived, and the results demonstrated that the accuracy of this model for the recognition of the geographical origin of rice varieties was 96.7%, which was higher than that of the Raman and mid-infrared models. [10].

The development of modern information technology has also driven the growth of the e-commerce field. More and more technologies are applied to the e-commerce field to promote the development of e-commerce. For example, Wang et al. proposed a recommendation model combined with collaborative filtering algorithm to improve the satisfaction of e-commerce website users. Through empirical analysis, the results showed that this model could accurately recommend e-commerce website users, greatly improving the stickiness of users [11]. Gao et al. proposed a dynamic intelligent hybrid recommendation algorithm based on DEMATEL to solve the problem that e-commerce platforms were difficult to retain customers. This algorithm could evaluate and rank customers' preferences and better help e-commerce platforms retain users [12]. To better establish the framework of the concept of e-commerce shopping experience, the Pentina et al. proposed a comprehensive measurement scale based on grounded theory. The results showed that this study could help retailers develop strategies to attract consumers more and improve consumers' shopping experience [13]. Saha and Sahney proposed a measurement model based on information search and social agent dimensions to accurately predict consumers' purchase intention on e-commerce platforms. The model was employed for empirical analysis. The findings indicated that the model could accurately identify the relationship between consumer information search dimensions, thereby assisting e-commerce platforms in effectively marketing products and enhancing their sales. [14]. An et al. proposed a classifier model based on data sets to estimate the gender of customers, aiming at the problem that the gender information of e-commerce articles recommendation system was difficult to identify correctly. The experimental results showed that the accuracy of this method was 78%, which could greatly increase the recommendation power of e-commerce articles recommendation system [15].

As evidenced by the preceding pertinent studies, Table 1 provides a comprehensive overview of the principal contributions, method, and outcomes of the relevant studies.

Table 1: Main contributions, method and results of relevant studies

Author	Major contribution	Method	Result
Gong et al. [6]	The time series data set of energy consumption of primary school buildings is mined	K-shape algorithm and Apriori algorithm were used for cluster analysis and association analysis	The curves obtained by data mining algorithm can effectively translate the running characteristics of primary school buildings
Gao Y et al. [7]	Tap the value of massive data in the Internet of Things	A collaborative filtering recommendation algorithm based on multi-information source fusion is proposed	The accuracy, recall rate and F1 value of recommendation results are better than other algorithms
Zare et al. [8]	Identify factors affecting hearing loss in industrial workers	Model based on neural network algorithm	The model helps workers better identify factors affecting hearing loss and provides data support for hearing protection programs
Becker et al. [9]	Improved capillary test results interpretation	Data mining algorithm was used to analyze the differences between capillary and venous blood technical analytes	Improved guidance on the interpretation of capillary test results in children
Sha et al. [10]	Improve the ability to identify the geographical origin of rice	A data mining and fusion model based on Raman spectrum and mid-infrared	The accuracy of geographic origin identification was 96.7%, higher than that of Raman and

		spectrum is proposed	mid-infrared models
Wang et al. [11]	Improve e-commerce website user satisfaction	Combined with collaborative filtering algorithm recommendation model	Precise recommendations to improve user engagement
Gao et al. [12]	Help e-commerce platforms retain customers	Dynamic intelligent hybrid recommendation algorithm based on DEMATEL	Evaluate and rank customer preferences to help e-commerce platforms retain users
Pentina et al. [13]	Establish a conceptual framework for e-commerce shopping experience	Comprehensive measurement scale based on grounded theory	Help retailers develop strategies that are more appealing to consumers and enhance the shopping experience
Saha et al. [14]	Study the dimension relationship of consumer information search	Measurement model based on information search and social agent dimension	Correctly find the dimensional relationship of information search to help e-commerce platforms with accurate marketing
An et al. [15]	Improve the gender identification accuracy of e-commerce recommendation system	Classifier model based on data set	The accuracy of gender identification is 78%, which improves the success rate of recommendation

In Table 1, extant research has encompassed the application of data mining and classification algorithms across a multitude of domains, including building energy consumption analysis, Internet of Things data value mining, health data analysis, agricultural product identification, e-commerce recommendation systems, and numerous others. Nevertheless, while these studies have yielded noteworthy outcomes in their respective domains, there is still scope for enhancement in the precision and efficacy of e-commerce big data processing. The research aims to optimize the data mining algorithm by combining the dimensional control mechanism and Spark mechanism, and improve the efficiency of e-commerce big data processing. At the same time, the accuracy of classification is improved by improving KNN classification algorithm. The aforementioned enhancements will not only furnish the e-commerce platform with more robust data-driven support for precision marketing, but will also facilitate the strategic transformation of the e-commerce platform. Therefore, the study aims to fill the gaps in the existing research on the accuracy and efficiency of e-commerce big data processing, so as to provide a data basis for the precision marketing of e-commerce customers and help the strategic transformation of e-commerce platforms.

3 Data classification and mining algorithms in e-commerce big data

3.1 Application of data mining algorithms in e-commerce

To help provide new methods and ideas for precision marketing in e-commerce, the research uses data mining algorithms and classification algorithms to collate and analyse data from e-commerce. Due to the wide range and complexity of e-commerce data sources, the research uses a combination of dimensional control mechanism and Spark mechanism to mine e-commerce data. The dimensional control mechanism is mainly used in the process of data mining to reduce the dimension of the data set in order to process the data more efficiently and extract useful information. In the context of e-commerce big data, dimensional control assumes particular importance due to the extensive range and intricate complexity of the data sources involved. The process of dimensional control mechanism is as follows: First, data is collected from each node of the e-commerce platform and integrated into a unified data set. The consolidated dataset is then thoroughly scanned to understand the amount of data in each data dimension. Then the node with the largest amount of data is selected as the initial mining node. Subsequently, the remaining nodes are reordered in accordance with the magnitude of the data volume. Only the nodes with the largest data volume are mined in each sorting cycle. Finally, according to the mining results, a data model tree is constructed to represent the relationship and importance between the data. The data is first scanned and dimensionally controlled by constructing a data model tree. This is done by scanning the data from the entire e-commerce network and placing the data obtained from each node in a unified dataset, which is represented by $J\langle 1,2,\cdots,n\rangle$. The node with the largest amount of data in the dataset J i is then used as the initial mining node and the other nodes are re-ranked. Only the node with the largest amount of data is mined in each sorting cycle, and the data model tree constructed for the study is shown in Figure 1.

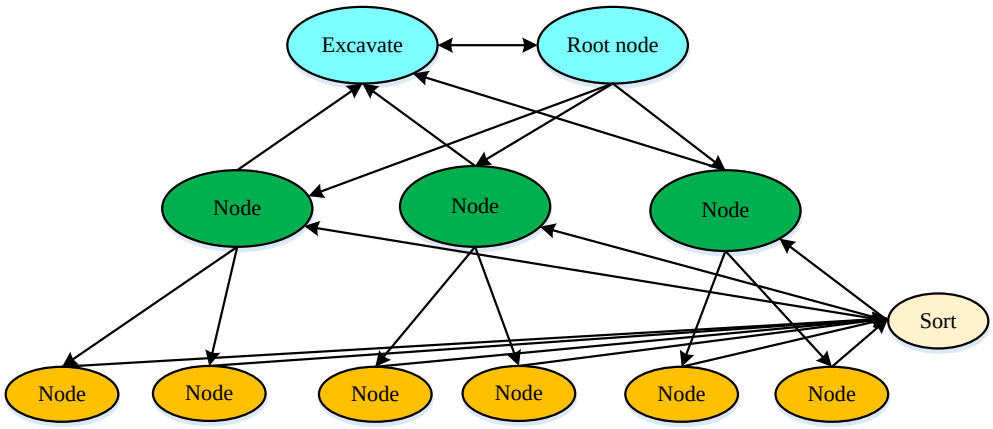


Figure 1: Data model tree

The data mining of the data model tree allows for the retrieval of data from the nodes with the largest amount of data. However, to achieve accurate marketing of e-commerce customers, it is necessary to supplement this data with information from other sources. It is also necessary to collect data on the consumer's consumption behavior. In the process of data formation, the relationship between the consumption behavior of e-commerce customers and the nodes is closer. However, the relationship between the two is not a positive fluctuating relationship, so it is necessary to build a user behavior mining tree based on the data model tree by making changes according to the consumption behavior and habits of e-commerce customers. The user behavior mining tree constructed by the study is shown in Figure 2.

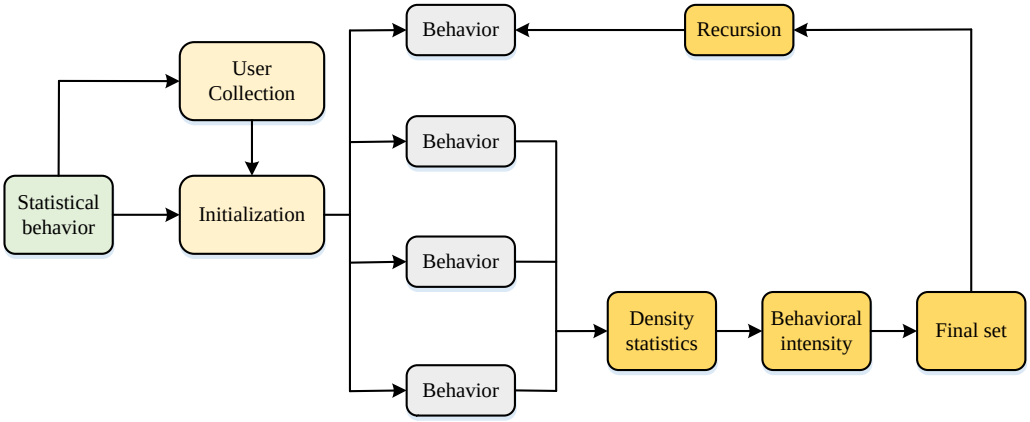


Figure 2: User behavior mining tree

As shown in Figure 2, the user behavior mining tree sorts the consumer's consumption behavior, viewing behavior, etc. according to the cycle, and then ranks the nodes with the highest user behavior ranking with their corresponding nodes with the largest amount of data. In this way, a more accurate user consumption behavior can be obtained. After forming two models of data model book and user behavior mining tree through dimensional control mechanism, in order to improve the data mining effect, the research uses Spark mechanism to enhance the correlation between data model book and user behavior mining tree. The specific process of Spark mechanism is shown in Figure 3.

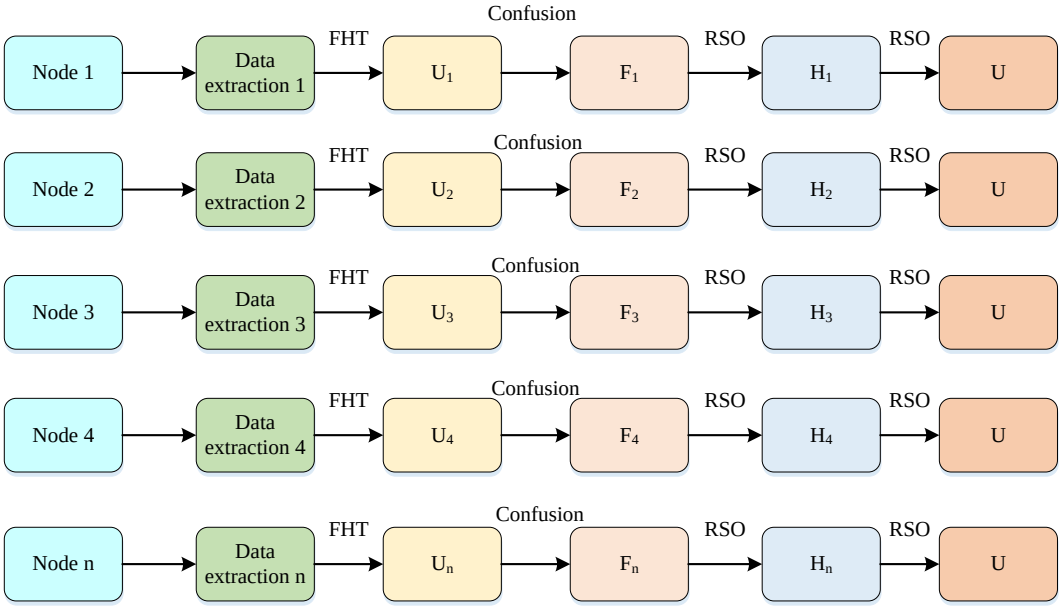


Figure 3: Spark mechanism flow chart between the two models

Figure 3 shows the flow diagram of the Spark mechanism between the data model tree and the user behavior mining tree, from which it can be seen that the Spark mechanism flow is divided into four steps. Firstly, the data set obtained from the full scan is sorted by column and transformed by Fourier formula. Secondly, the coupling operation between the data model tree and the user behavior mining tree is carried out to generate the coupling sequence. Subsequently, random sequence alterations are introduced to the coupled sequence with the objective of enhancing the diversity of the data set. Finally, the RSO mechanism is used to obfuscate the sequence and generate the final output sequence. To be specific, the first step is to sort the data set obtained through a full scan $J\langle 1,2,\cdots,n\rangle$ by column. The data sorted by column is transformed using a Fourier formula, as shown in equation (1).

$$F(\omega) = \square\left[f(t)\right] = \int_{-\infty}^{\infty} f(t)e^{-iwt}dt \qquad (1)$$

In equation (1) $F(\omega)$ is the bishop function of $f(t)$ and $f(t)$ is called the bishop function of $F(\omega)$. In addition to the Fourier transform of the data set $J\langle 1,2,\cdots,n\rangle$, the second step of the process couples the data model tree and the user behavior mining tree, and defines the coupled sequence as the sequence $U_1,U_2,U_3,\cdots U_n$. In the coupling process, a Spark mapping is involved, and the mapping formula is shown in equation (2).

$$FHT(s,t) = \int \beta(s)\tan\left[\frac{1}{M}s(x+\frac{\pi}{M})\right]ds\int \beta(t)\omega t\left[\frac{1}{M}t(y+\frac{\pi}{M})\right]dt \qquad (2)$$

In equation (2), s is the data in the data model tree and t is the data in the user behavior mining tree. x represents the key data in the data model tree, y represents the key data in the data model tree, and M represents the total amount of data. The third step of the Spark mechanism process is to perform a random sequence change on the sequence $U_1,U_2,U_3,\cdots U_n$ resulting from the coupling operation to obtain a new sequence $F_1,F_2,F_3,\cdots F_n$. The next step of the Spark process is to perform an RSO obfuscation operation on the sequences $F_1,F_2,F_3,\cdots F_n$ and $U_1,U_2,U_3,\cdots U_n$ using the RSO mechanism to obtain the obfuscated sequence $H_1,H_2,H_3,\cdots H_n$. The final step is to perform another RSO obfuscation of the obfuscated sequence $H_1,H_2,H_3,\cdots H_n$ to obtain the final output sequence U. The final output sequence U is a collection of data mined by the data mining algorithm. After obtaining this collection of sequences, the research uses a classification algorithm to accurately classify the data in order to predict the consumer's consumption intention and thus help the e-commerce platform to accurately market to its customers.

3.2 Improved KNN algorithm based on k-value selection strategy

The K-Nearest Neighbor (KNN) classification algorithm is a common classification algorithm, which is widely used in many classification situations because of its simplicity and high classification accuracy [16]. The distance metric of the algorithm has a great influence on the classification accuracy of the algorithm [17]. In this algorithm, Min's distance is generally used as the distance measure. Its formula in the feature space is equation (3).

$$d_p(x_i,x_j) = \left(\sum_{l=1}^{n}\left|x_i^{(l)} - x_j^{(l)}\right|^p\right)^{\frac{1}{p}} \qquad (3)$$

In equation (3), x_i is the eigenvector of i. n represents the number of eigenvalues of the eigenvector. p represents different distance metric patterns. When

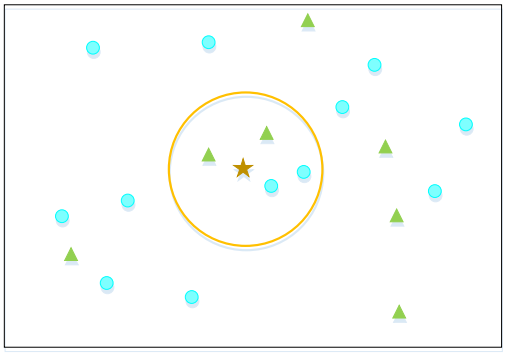
p is 1, the distance represented by the formula is the Manhattan distance, then the formula is shown in equation (4).

$$d_1(x_i, x_j) = \sum_{l=1}^{n} \left| x_i^{(l)} - x_j^{(l)} \right| \tag{4}$$

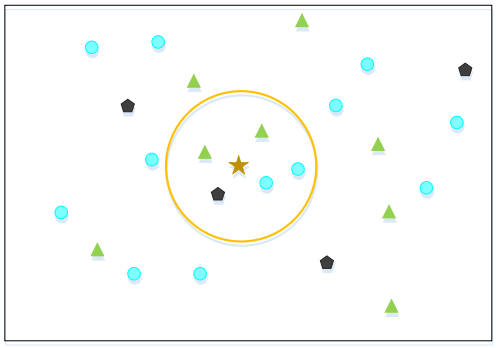
When p is 2, the distance expressed by this formula is called the Euclidean distance. The formula is shown in equation (5).

$$d_2(x_i, x_j) = \left(\sum_{l=1}^{n} \left| x_i^{(l)} - x_j^{(l)} \right|^2 \right)^{\frac{1}{2}} \tag{5}$$

In the KNN algorithm, different values of p will result in different nearest neighbors being selected, so the choice of the appropriate distance algorithm is quite crucial for the KNN algorithm. In addition to the choice of distance brightness, the choice of K value is also very important in the KNN algorithm. If the K value is too small, the classification model will become complex and the overall accuracy of the classification model will be reduced. If the K value is excessive, samples that are not sufficiently similar to the samples to be tested will also be included in the classification process, resulting in a reduction in the accuracy of the classification. In the traditional KNN classification algorithm, the sample to be tested is classified by comparing the number of each kind of each of the K neighboring samples of the sample to be tested. Moreover, the sample to be tested is classified into the class with the highest number of the K neighboring samples [18]. However, the traditional KNN algorithm is prone to misclassification due to the fact that the class with the highest number of K neighbour occurrences is tied for first. Figure 4 shows the classification error of traditional KNN algorithm.



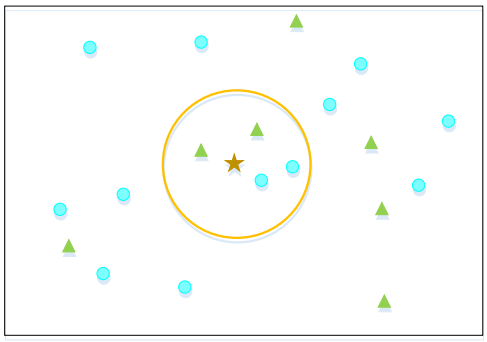
(a) There are exactly two types of samples with the same number of K adjacent samples



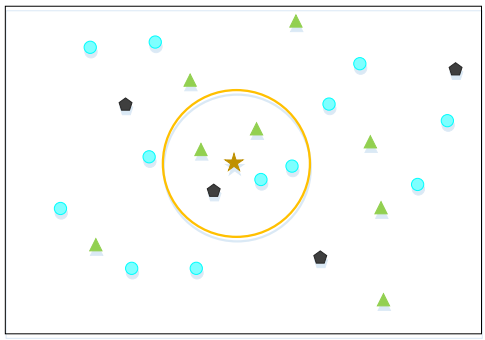
(b) There are three types of K adjacent samples

Figure 4: Example of Traditional KNN Algorithm

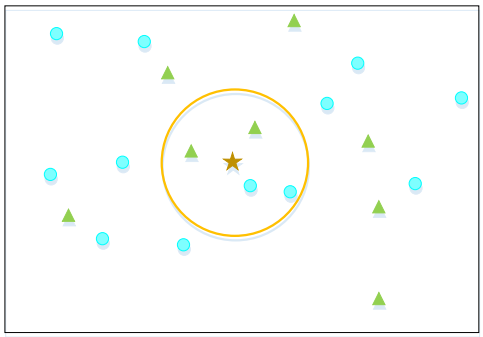
As shown in Figure 4(a), the number of both triangles and circles in the proximity samples of the dataset is 2, resulting in the inability to determine whether the samples to be tested are classified as triangles or circles. Furthermore, the conventional KNN classification algorithm typically comprises an odd number of K values. While this configuration is less susceptible to classification errors in binary classification problems, it can nevertheless result in classification errors in multi-classification problems. [19]. As shown in Figure 4(b), the K value is 5, and the five selected neighboring samples are two triangles, two circles and one pentagon. At this point, the number of triangles and circles are tied for first place, and the classification of the samples to be tested still cannot be accurately judged. To improve the above problem, the study proposes an improved algorithm based on the K-value selection strategy, in which the K-value can be taken as an even number. When there is an equal number of samples among the K neighboring samples, the average distance from the sample to be tested is used as the classification index. Moreover, the sample to be tested is classified as the class with the average distance from the sample to be tested. Figure 5 shows the classification of improved KNN algorithm.



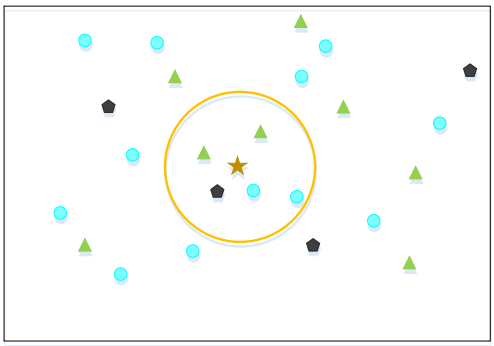
(a) There are exactly two types of samples with the same number of K adjacent samples



(b) There are three types of K adjacent samples



(c) The average distance between two types of selected adjacent samples is equal (second classification)



(d) The average distance between two types of selected adjacent samples is equal (multi classification)

Figure 5: Example of improved KNN algorithm

In Figure 5(a) and (b), the number of triangular samples and circular samples are equal, but the average distance of triangles from the sample to be tested is short, so the sample to be tested is classified as triangular. In Figure 5(c) and (d), the number of triangular samples and circular samples are equal, and the average distance of both from the sample to be tested is equal. Therefore, the sample to be tested is classified as whichever category is the closest sample to the sample to be tested. Subsequently, in Figure 5(c) and (d), the sample to be tested should be classified as a sample of the circle class which is closer to it. The principal flow chart of the optimal classification algorithm based on the K-value selection strategy is shown in Figure 6.

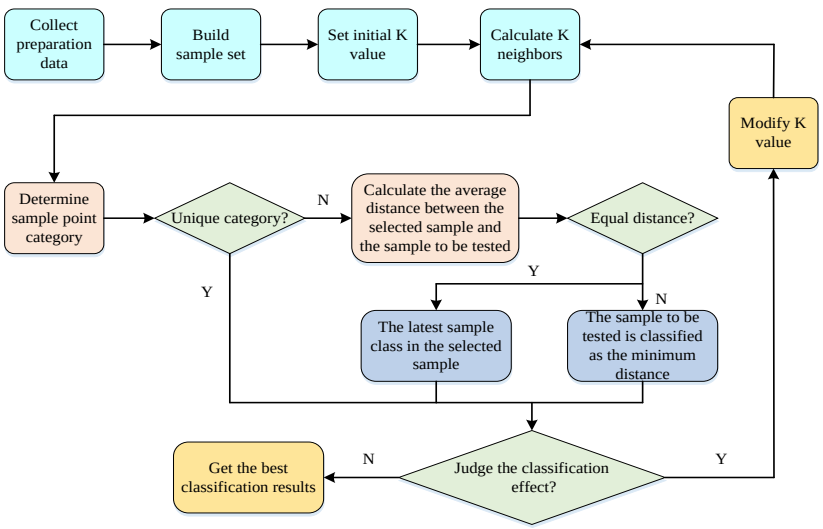


Figure 6: Flow chart of optimized KNN algorithm

As shown in Figure 6, the optimized KNN algorithm process is divided into nine steps. The first step is to collect and prepare the data from the database and extract the features of the data. The second step is to use the data in the dataset as the training dataset and construct the sample set X. The range of the set X is shown in equation (6).

$$x_i \in X \subseteq R^n \tag{6}$$

In equation (6), x_i denotes the i th sample, while R^n denotes the n dimensional space. The next step is to make an initial setting of the value of K. The fourth step is to use the Euclidean distance formula to find the nearest K samples in the set X to the sample to be tested x, which is denoted as $N_k(x)$. At this point, the Euclidean distance between the samples x_i and x_j is shown in equation (7).

$$d(x_i, x_j) = (\sum_{l=1}^{n} (x_l^i - x_l^j)^2)^{\frac{1}{2}} \tag{7}$$

In equation (7), x_l^i is the l attribute of the i sample. The category label of x_i is then set to y_i. The range of y_i is given in equation (8).

$$y_i \in Y = \{c_1, c_2, c_3 \cdots, c_k\} \tag{8}$$

In equation (8), k represents the number of categories in the training sample set, while c_i represents the label of the i category. The fifth step is to determine the formula for the category y of the sample x to be tested according to the classification rules as shown in equation (9).

$$\begin{aligned} &y = \\ &\arg\max_{c_j} \sum_{y_i} I(y_i = c_j), \\ &i = 1, 2, 3 \cdots, N; \; j = 1, 2, 3 \cdots, K \end{aligned} \tag{9}$$

In equation (9), I represents the feature function. The sixth step of the optimization KNN algorithm is to determine whether the obtained y value is unique. If the y value is unique, the classification effect is judged. If the classification effect is good, a conclusion is drawn. If the classification effect is not good, the K value is reset in the third step until the classification effect is good. If it is found that the value of y is not unique, it means that there are two types of samples with the same number of samples among the K close samples. Then the average distance between the two types of samples with the same number of samples needs to be calculated, and the formula is shown in equation (10).

$$\begin{cases} d_{y_1} = \dfrac{1}{m} \sum_{1}^{m} d(x_i \in c_a, x) \\ d_{y_2} = \dfrac{1}{m} \sum_{1}^{m} d(x_i \in c_b, x) \end{cases} \tag{10}$$

In equation (10), c_a and c_b represent the two types of samples with the same number of samples and m represents the number of samples in both types. d indicates the distance between each sample and the sample center. After finding the average distance of the two types of samples, the average distance of the two types of samples d_{y_1} and d_{y_2} are compared. In the event of equality, the result is equation (11).

$$y = c_j, \; j \in \{a, b\} \tag{11}$$

In equation (11), c_j is determined by equation (12).

$$c_j \to \min_{x_i \in N_K(x)} (d(x_i \in c_a, x), d(x_i \in c_b, x)) \tag{12}$$

In equation (12), c_j is the class to which the nearest of the K proximate samples x_i belongs, among the samples to be tested x. If d_{y_1} and d_{y_2} are not equal, then equation (13) is given.

$$y = if, (d_{y_1} < d_{y_2}, c_a, c_b) \tag{13}$$

Equation (13) indicates that the sample to be tested is classified as a sample class with a small distance, at which point the classification effect is evaluated to determine whether it has reached a high level. If so, the result is output. If not, the process resumes with the third step, wherein the K value is reset to continue the cycle.

3.3 Application of Improved KNN algorithm in e-commerce

After deep mining of data in e-commerce using the dimensional control mechanism and Spark mechanism, the data in the mined data sequence U needs to be classified to better help e-commerce platforms in precision marketing. The study uses the improved KNN algorithm based on K-value selection to classify the mined data U, and the classification results can be used to analyse and predict the corresponding data of consumers. The flow of classification using the improved KNN algorithm is shown in Figure 7.

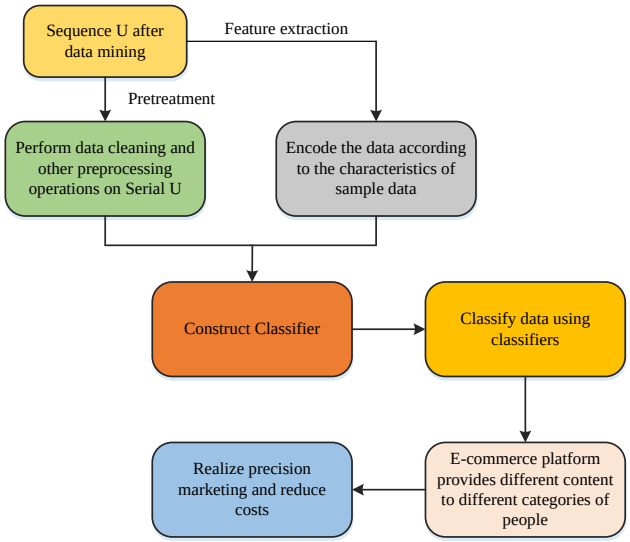


Figure 7: Application flow chart of improved KNN algorithm in e-commerce data classification

As shown in Figure 7, the application of the improved KNN classification algorithm in e-commerce data classification is mainly divided into several steps, such as data pre-processing, data feature coding, constructing classifiers and data classification. Among them, the data pre-processing step is mainly to process the U obtained from data mining. The pre-processing is mainly to classify the personal elements such as consumption habits and consumption behaviors of e-commerce users in multiple dimensions. Moreover, the people who use the e-commerce platform less are eliminated so as not to affect the accuracy of predicting user behavior. The data feature coding is the more important part of the whole classification process, in which the distance between the sample data will be measured by means of feature coding. Since the factors that affect online shopping are mainly gender, age and geography, the study focuses on feature coding from these three parts [20]. The KNN classifier is constructed by different feature coding, and a suitable K value is derived by training the data in this step. Finally, consumers whose distance is within the K value are classified into the same class of consumers to achieve accurate classification. Additionally, the same message is delivered to this class of consumers to reduce the cost in the process of accurate marketing, thus realizing the strategic transformation of the e-commerce platform.

4 Performance comparison of data mining and improved KNN algorithms

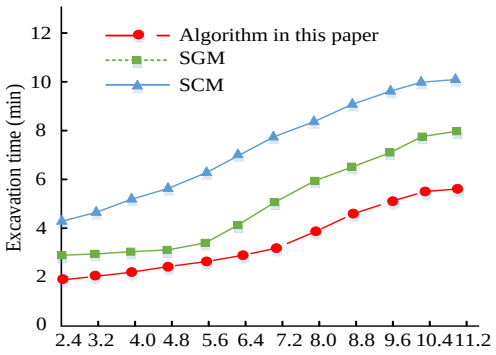
4.1 Data mining algorithm performance comparison experiments

To test the performance of the data algorithms proposed in the study, the mining algorithms proposed in the study are simulated and tested using Matlab. The results are compared and analyzed with the common SCM mining algorithm and SGM mining algorithm. To ensure the accuracy and repeatability of the experimental results, a high-performance hardware environment is used, including the Intel Core i9-9900K processor (3.6 GHz, 3.6 GHZ; 8 cores, 16 threads), 64 GB DDR4 RAM, 1 TB NVMe SSD, and NVIDIA GeForce RTX 2080 Ti graphics card. On the software side, the study is based on Windows 10 Pro 64-bit operating system using MATLAB R2021a programming environment. The study also combines MATLAB built-in functions and custom scripts with specially written data mining algorithms to fully support data processing and analysis. In the process of comparison experiment, two sets of different simulation parameters are used to test the three algorithms, and the different simulation parameters of the two sets are shown in Table 2. The selection of simulation parameters is based on the common scenarios of e-commerce big data processing.

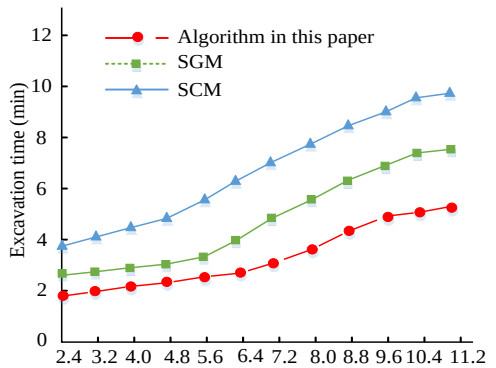
Table 2: Two groups of different simulation parameters

Simulation parameter 1		Simulation parameter 2	
Parameter	Value	Parameter	Value
Data mining time/min	20	Data mining time/min	25
Excavation speed/(Kb/s)	>100	Excavation speed/(Kb/s)	>120
Number of network nodes	>1200000	Number of network nodes	>1500000
Number of e-commerce users	>1200000	Number of e-commerce users	>1500000

In Table 1, the mining time and mining intensity serve to reflect the processing speed of the algorithm in practical applications, while the number of network nodes and the number of electric business households are employed as a means of simulating the data scale in a big data environment. The selection of these parameters aims to comprehensively evaluate the performance of the algorithm under different conditions. Under the two simulation parameters shown in Table 2, the research tested the mining time and mining error rate of the three algorithms under different mining intensity. Moreover, the tested data set is a large e-commerce data set generated by simulation to reflect the actual data situation of the e-commerce platform. The comparison of the mining times of the three algorithms is shown in Figure 8.



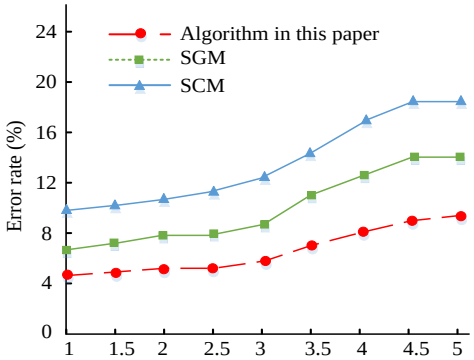
(a) Mining time test of three algorithms under the first parameter



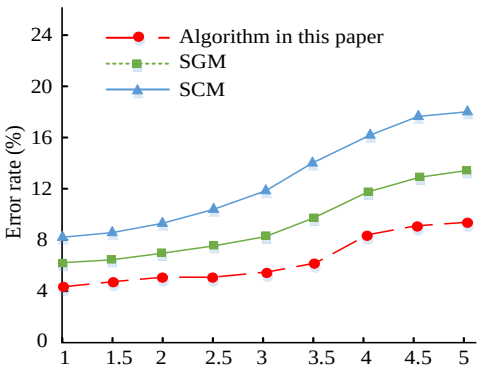
(b) Mining time test of three algorithms under the second parameter

Figure 8: Comparison of mining time of three mining algorithms under different mining intensities

Figure 8 shows the comparison of the mining time of the three algorithms under simulation parameter 1. In Figure 8(a), under simulation parameter 1, the mining time of the proposed data mining algorithm is the lowest at each mining intensity. Moreover, it reaches the maximum at the mining intensity of 11.2Kb/s, and the mining time is 4.8min. Figure 8(b) shows the comparison of the mining time of the three mining algorithms under simulation parameter 2. The mining time of the proposed data mining algorithm is also lower than the other two algorithms at all mining intensities under simulation parameter 2. Moreover, it is maximum at the mining intensity of 11.2 Kb/s, and the mining time is 4.6 min at this time. The comparison of the mining error rates of the three algorithms under the two simulation parameters is shown in Figure 9.



(a) Mining time test of three algorithms under the first parameter



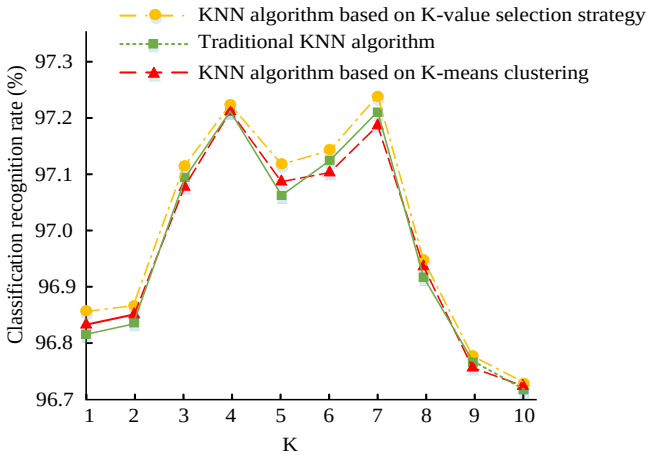
(b) Mining time test of three algorithms under the second parameter

Figure 9 Comparison of mining error rates of three algorithms under different parameters

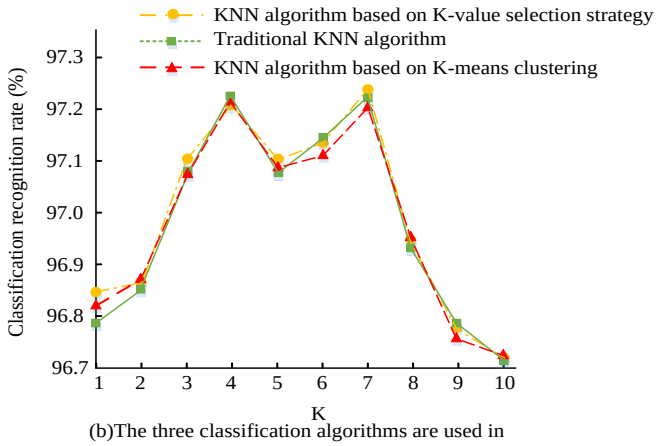
Figure 9 depicts the comparative performance of the data mining algorithms, namely the SCM, SGM, and the proposed algorithm, in terms of the mining error rate under two distinct simulation parameters. Among them, Figure 9(a) shows the comparison curves of the mining error rates corresponding to different mining intensities of the three mining algorithms under simulation parameter 1. The mining error rate of the SCM mining algorithm is the highest among the three algorithms, with the lowest mining error rate of 9.8%, which is much higher than that of the proposed algorithm. Figure 9(b) shows a comparison of the mining error rates of the three mining algorithms for different mining intensities under simulation parameter 2. The SCM mining algorithm has the highest mining error rate among the three algorithms, with the lowest mining error rate of 8.0%, which is also much higher than that of the proposed algorithm. The above findings indicate that the performance of the proposed data mining algorithm is better than that of general data mining algorithms, and that the mining accuracy of e-commerce big data can be improved by using this algorithm.

4.2 Data classification algorithm performance comparison experiments

To validate the actual classification performance of the improved KNN algorithm, the Mnist dataset is used in the study. This dataset is a standard dataset widely used in the field of machine learning, including handwritten digital images, and is suitable for validating and improving the classification algorithm. Therefore, this study uses selected Mnist dataset to conduct comparative experiments on KNN algorithm based on K-value selection strategy, traditional KNN algorithm and KNN algorithm based on K-means clustering. The Mnist dataset contains 60,000 training sets and 10,000 test sets, with the training set divided into 10 classes and each class divided into 400 subclasses. The Mnist dataset is used to compare the three algorithms twice. In the experiments, different K values are chosen to compare the classification accuracy and classification time of the three different algorithms. The classification recognition rates of the three algorithms at different K values are shown in Figure 10.



(a) The three classification algorithms are used in MNIST for the first time Classification recognition rate on data set



(b)The three classification algorithms are used in MNIST for the second time Classification recognition rate on data set

Figure 10: Classification recognition rate of three algorithms under different K values

The classification recognition rate curves of the three algorithms under different K values are shown in Figure 10. Figure 10(a) shows the classification recognition rate curves of the three algorithms on the dataset for the first time. The classification recognition rates of the three classification algorithms are generally not very different, while the KNN improvement algorithm based on the K-value selection strategy has a slightly higher classification recognition rate than the other two classification algorithms. Moreover, it has a maximum value of 97.23% when K=7. Figure 10(b) shows the classification recognition rate curves of the three algorithms for the second time on the dataset. The recognition accuracy of the three algorithms is not very different, and the classification recognition rate of the KNN improvement algorithm based on the K value selection strategy is slightly higher than the other two algorithms among the three algorithms. Moreover, it has a maximum value of 97.24% when K=7. The classification times of the three algorithms at different K values are shown in Table 3.

Table 3: Classification time of three algorithms under different K values

/	The first comparative experiment			The second comparative experiment		
K	KNN	KNN based on K-value selection strategy	KNN based on K-means clustering	KNN	KNN based on K-value selection strategy	KNN based on K-means clustering
1	403	29.5	31.5	408	28.9	30.9
2	398	29.6	32.3	402	28.6	31.8
3	408	28.3	30.5	406	27.8	30.5
4	405	27.6	31.8	401	27.3	30.7
5	411	28.8	30.9	395	28.5	31.3
6	388	29.7	31.4	413	29.7	32.1
7	403	30.2	32.3	389	28.9	31.4
8	402	29.4	30.5	397	27.8	30.9
9	395	30.5	31.6	403	28.5	31.4
10	401	27.8	30.8	412	29.1	30.7

Table 3 shows the classification times of the three classification algorithms under different K values. In the initial comparison experiment, the classification time of the traditional KNN algorithm is considerably longer than that of the other two algorithms, with a classification time of approximately 400 seconds. In contrast, the classification time of the KNN algorithm based on the K-value selection strategy is notably shorter than that of the other two algorithms, with the lowest value of 27.6 seconds when K is taken as 4. In the second comparison experiment, the KNN algorithm based on the K-value selection strategy has a lower classification time than the other two classification algorithms for all K values, with a minimum classification time of 27.3 seconds at K of 4. These results show that the KNN algorithm based on the K-value selection strategy outperforms the other two classification algorithms in terms of classification recognition rate and classification time. Therefore, the classification of e-commerce data can be better achieved by using this classification algorithm. In addition, in order to further analyze the superiority of the improved KNN algorithm proposed in this study, the study also compares it with the current relatively new Gradient Boosting Decision Tree (GBDT) and optimized random forest, respectively. ORF and Light Gradient Boosting Machine (LightGBM) classification algorithm are compared, and ROC curve, F1 value and accurate rate equivalent are used as comparison indexes. ROC curves of four algorithms are shown in Figure 10.

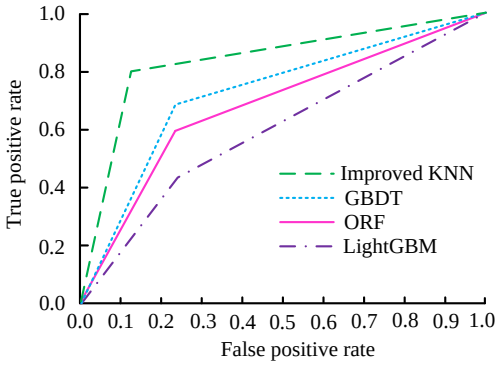


Figure 11: Comparison results of ROC curves of the four algorithms

In Figure 10, the area under ROC curve of the improved KNN algorithm proposed in this study is 0.88, which is higher than the 0.76 of GBDT algorithm, 0.66 of ORF algorithm and 0.61 of LightGBM algorithm. The aforementioned results demonstrate that the enhanced KNN algorithm exhibits superior performance in comparison to the three benchmark algorithms, as evidenced by the ROC curve analysis. In addition, the accuracy rate, classification error, F1 value, recall rate and other indicators of the four algorithms are statistically obtained in Table 4.

Table 4: Performance comparison indexes of the four classification algorithms

Algorithm type	Precision	Classification Error	F1-Score	Recall
Improved KNN algorithm	0.92	0.08	0.91	0.93
GBDT	0.85	0.15	0.87	0.88
ORF	0.78	0.22	0.79	0.80
LightGBM	0.75	0.25	0.77	0.76

According to the analysis of Table 4, the accuracy rate of the improved KNN algorithm reaches 0.92, which is significantly higher than GBDT's 0.85, ORF's 0.78 and LightGBM's 0.75. It indicates that the improved KNN is more accurate in identifying positive samples. At the same time, the classification error is only 0.08, which is much lower than other algorithms, indicating that the misjudgment rate is very low. The F1 value is as high as 0.91, which proves that the algorithm achieves a good balance between accuracy and recall rate. The recall rate of 0.93 indicates that the majority of positive samples can be accurately identified, thereby further substantiating the efficacy of the enhanced KNN algorithm. In conclusion, the classification performance of the improved KNN algorithm is significantly better than that of the comparison algorithm.

5 Discussion

This study proposed an optimized data mining algorithm and an improved KNN classification algorithm for e-commerce big data processing, and demonstrated their superior performance through simulation experiments. Compared with the existing research, the study showed significant advantages in several aspects. In terms of data mining algorithms, Gong et al. used k-shape algorithm and Apriori algorithm to analyze primary school building energy consumption data. Although effective results was obtained, these traditional algorithms were limited in efficiency and accuracy due to the complexity and scale of e-commerce big data. Similarly, the collaborative filtering recommendation algorithm based on the fusion of multiple information sources proposed by the Gao team improved the value mining ability of iot data, but it might be difficult to achieve ideal performance when processing high-speed dynamic data in the field of e-commerce. By combining the dimensional control mechanism and Spark mechanism, this study significantly improved the efficiency and accuracy of data mining algorithms in e-commerce big data processing, thus making up for the shortcomings of existing studies. In terms of classification algorithms, although existing studies such as An team used classifier models to estimate gender information in e-commerce recommendation systems and achieved certain results, there was still room for improvement in its accuracy. The proposed method entailed an enhancement of the KNN algorithm and the incorporation of a K-value selection

strategy. This approach not only elevated the classification recognition rate but also markedly curtailed the classification time, which is a pivotal consideration in the real-time processing of e-commerce data. In addition, compared with advanced classification algorithms such as GBDT, ORF and LightGBM, the study found that the improved KNN algorithm showed better performance in multiple indicators such as ROC curve, accuracy rate, F1 value and recall rate, which further verified its effectiveness and efficiency in e-commerce big data classification.

In conclusion, the research proposes an optimized data mining algorithm and an improved KNN classification algorithm, effectively solving the efficiency and accuracy issues in e-commerce big data processing and providing robust data support for precision marketing of e-commerce platforms. Compared with existing studies, this study has shown significant advantages and innovation in algorithm design, experimental verification and result analysis, and has made important contributions to the development of e-commerce big data processing.

6 Conclusion

With the popularity of online shopping, more and more people are using e-commerce platforms. The growth in the number of e-commerce platform users has led to a decrease in the accuracy of the data mining and classification algorithms applied to them. To enhance the efficacy of this approach, the research integrates data mining methods and data classification algorithms, employing a dimensional control mechanism and the Spark mechanism to mine data from e-commerce platforms, and subsequently utilizing an enhanced KNN algorithm to accurately categorize the mined data. The classified data can then be leveraged for precise marketing initiatives targeting e-commerce platform users. The results showed that the mining time of the proposed data mining algorithm was 4.6min, which was lower than the 7.8min of SGM algorithm and 10.1min of SCM algorithm. Moreover, the mining error rate was 4.2%, which was lower than the 6.5% of SGM algorithm and 9.8% of SCM algorithm. The classification recognition rate and classification time of the KNN algorithm based on the K-value selection strategy were 97.23% and 27.3 seconds, respectively, which were both better than the traditional KNN algorithm and the KNN algorithm based on K-mean clustering. The above results indicate that both the data mining algorithm proposed in the study and the KNN algorithm based on the K-value selection strategy outperform other similar algorithms. While these two algorithms demonstrate superior performance in the study, the selected dataset is more limited in scope. Consequently, subsequent research will focus on enhancing the algorithms' capabilities in real-world datasets.

References

[1] S. Song, W. Peng, and Y. Zeng, “Optimal add-on items recommendation service strength strategy for e-commerce platform with full-reduction-promotion,” RAIRO-Operations Research, vol. 56, no. 2, pp. 1031-1049, 2022. https://doi.org/10.1051/ro/2022037

[2] E. T. Group, “Take your packaged gas offerings online: How general distributing company transformed its e-commerce platform with EvolutionX,” Gasworld: Incorporating CryoGas International, vol. 59, no. 8, pp. 20-21, 2021.

[3] J. K. Sarmah, and S. Baruah, “A comparative study on use of data mining algorithm in EDM development,” IOSR Journal of Computer Engineering, vol. 23, no. 1, pp. 44-48, 2021. https://doi.org/10.9790/0661-2301014448

[4] X. Liu, and Q. Zhou, “Intelligent manufacturing system based on data mining algorithm,” International Journal of Grid and Utility Computing, vol. 12, no. 4, pp. 396-405, 2021. https://doi.org/10.1504/ijguc.2021.119554

[5] P. Wang, and N. Zhang, “Decision tree classification algorithm for non-equilibrium data set based on random forests,” Journal of Intelligent & Fuzzy Systems, vol. 39, no. 2, pp. 1639-1648, 2020. https://doi.org/10.3233/jifs-179937

[6] Q. Gong, X. Liu, Y. Zeng, and S. Han, “An energy efficiency solution based on time series data mining algorithm on elementary school building,” International Journal of Low-Carbon Technologies, vol. 17, pp. 356-372, 2022. https://doi.org/10.1093/ijlct/ctac008

[7] K. Huang, J. Yuan, Z. Zhou, and X. Zheng, “Analysis and evaluation of heat source data of large-scale heating system based on descriptive data mining techniques,” Energy, vol. 251, no. 15, pp. 251-269, 2022. https://doi.org/10.1016/j.energy.2022.123834

[8] S. Zare, M. R. Ghotbiravandi, H. Elahishirvan, M. G. Ahsaeed, M. Rostami, and R. Esmaeili, “Modeling and predicting the changes in hearing loss of workers with the use of a neural network data mining algorithm: A field study,” Archives of Acoustics, vol. 45, no. 2, pp. 303-311, 2020. https://doi.org/10.24425/aoa.2020.133150

[9] M. Becker, T. Gscheidmeier, H. J. Groß, H. Cario, J. Woelfle, M. Rauh, and J. Zierk, “Differences between capillary and venous blood counts in children-A data mining approach,” International Journal of Laboratory Hematology, vol. 44, no. 4, pp. 729-737, 2022. https://doi.org/10.1111/ijlh.13846

[10] M. Sha, Z. Zhang, Y. Huang, M. Jiang, J. Liu, D. Gui, and P. Li, “Enhanced raman and mid-infrared spectroscopic discrimination of geographical origin of rice,” Data Mining and Data Fusion, vol. 36, no. 3, pp. 34-38, 2021.

[11] Y. Wang, Y. Zhu, Z. Zhang, H. Liu, and P. Guo, “Design of hybrid recommendation algorithm in

online shopping system," Journal of New Media, vol. 3, no. 4, pp. 119-128, 2021. https://doi.org/10.32604/jnm.2021.016655

[12] Y. Gao, H. Liang, and B. Sun, "Dynamic network intelligent hybrid recommendation algorithm and its application in online shopping platform," Journal of Intelligent & Fuzzy Systems, vol. 40, no. 5, pp. 9173-9185, 2021. https://doi.org/10.3233/jifs-201579

[13] I. Pentina, M. Zolfagharian, and A. Michaud-Trevinal, "Toward a comprehensive scale of online shopping experiences: a mixed-method approach," Internet Research, vol. 32, no. 3, pp. 814-842, 2022. https://doi.org/10.1108/intr-03-2021-0170

[14] M. Saha, and S. Sahney, "Exploring the relationships between socialization agents, social media communication, online shopping experience, and pre-purchase search: a moderated model," Internet Research, vol. 32, no. 2, pp. 536-567, 2022. https://doi.org/10.1108/intr-08-2020-0472

[15] Y. An, S. Meng, and H. Wu, "Discover customers' gender from online shopping behavior," IEEE Access, vol. 10, pp. 13954-13965, 2022. https://doi.org/10.1109/ACCESS.2022.3147447

[16] S. Zhu, C. He, M. Song, and L. Li, "Two-parameter KNN algorithm and its application in recognition of brand rice," Journal of Intelligent & Fuzzy Systems, vol. 41, no. 1, pp. 1837-1843, 2021. https://doi.org/10.3233/jifs-210584

[17] S. Wu, "Simulation of classroom student behavior recognition based on PSO-kNN algorithm and emotional image processing," Journal of Intelligent & Fuzzy Systems, vol. 40, no. 4, pp. 7273-7283, 2021. https://doi.org/10.3233/jifs-189553

[18] Z. Lei, L. Zhu, Y. Fang, X. Li, and B. Liu, "Anomaly detection of bridge health monitoring data based on KNN algorithm," Journal of Intelligent & Fuzzy Systems, vol. 39, no. 4, pp. 5243-5252, 2020. https://doi.org/10.3233/jifs-189009

[19] L. Han, Z. Su, and J. Lin, "A Hybrid KNN algorithm with Sugeno measure for the personal credit reference system in China," Journal of Intelligent & Fuzzy Systems, vol. 39, no. 5, pp. 6993-7004, 2020. https://doi.org/10.3233/jifs-200191

[20] P. Group, "Creating positive online shopping experiences," Pharmaceutical & Cosmetic Review, vol. 48, no. 4, pp. 12-13, 2021.