A Data Mining and Processing Method for E-Commerce Potential Customers Based on Apriori Association Rules Algorithm

Xian Zhou[1(✉)] and Hai Huang[2]

[1] Institute of Economics and Management, Zhi Xing College of Hubei University, Wuhan 430011, China
zhou_xian2001@163.com

[2] Shanghai Dong Hai Vocational and Technical College, Shanghai 200241, China

Abstract. In order to improve the effectiveness of e-commerce potential customer data mining and processing, a method based on Apriori association rule algorithm for e-commerce potential customer data mining and processing is proposed. Innovatively adopting a multidimensional tree structure to improve the Apriori association rule algorithm, using frequent itemsets as candidate itemsets, and further expanding on this basis by adding judgment conditions to reduce the frequency of scanning the database; The Vector space model is used to calculate the similarity between e-commerce potential customers, and the similarity is used as a scalar value to complete the accurate calculation. The e-commerce potential customers at different levels in customer transaction data are divided. Obtain a sticky evaluation system for potential e-commerce customers from the perspectives of perceived usefulness, perceived ease of use, perceived service, perceived security, and perceived interest, as the basic indicators for subsequent mining and processing. The Quicksort method is used to sort each data dimension in the e-commerce customer data set, and the improved Apriori association rule algorithm is used to realize data mining and processing of e-commerce potential customers through high-density grid. The experimental results demonstrate that the method innovatively utilizes the improved Apriori association rule algorithm to mine three types of customer behavior data with an accuracy of over 80%, which is in line with the actual situation. It improves the effectiveness of e-commerce potential customer data mining and processing, effectively mining e-commerce potential customers, and providing good basic data for e-commerce platforms to adjust marketing strategies.

Keywords: Apriori Algorithm · E-Commerce Potential Customers · Data Mining

1 Introduction

At present, China's e-commerce is booming, with various e-commerce platforms constantly emerging, while explosive data growth is occurring. There is valuable information hidden in this data that needs to be mined, such as e-commerce potential customer data.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
L. Yun et al. (Eds.): ADHIP 2023, LNICST 548, pp. 190–202, 2024.
https://doi.org/10.1007/978-3-031-50546-1_13

Data mining is the discovery of knowledge. Mining e-commerce potential customer data from the accumulated transaction data of e-commerce can find targeted and effective customer information in the massive information, which has important strategic significance for the development of modern enterprises [1–3]. E-commerce is a product of the rapid development of the internet industry, and it is a new type of business operation model. In an open online environment, buyers and sellers can engage in product selection, payment, and other trading activities on the Internet without meeting.

According to the survey report, e-commerce can be roughly divided into four types according to transaction partners: Firstly, B2C (Business to Customer): refers to e-commerce activities between enterprises and consumers, where consumers directly engage in commodity trading activities on the Internet. Secondly, B2B (Business to Business): refers to the business activities between enterprises. Can companies find suitable partners on the Internet? M line transaction activity. Thirdly, C2C (Consumer to Consumer): refers to e-commerce activities between consumers. Consumers can engage in sales activities in the form of retail investors on e-commerce platforms, such as Xianyu on the Alibaba platform. Fourthly, C2B (customer to business): refers to the business activities between consumers and enterprises. Consumers targeting the same consumer goods can purchase from enterprises in the form of teams, generate precise orders and produce goods, but this model is not yet mature.

Data mining can first provide customers with more comprehensive personalized services. By conducting in-depth mining and analysis of customer access information data, we can obtain customer purchasing behavior characteristics and preferences, understand customer habits, interests, potential needs, and loyalty issues to profile customer characteristics and reduce unnecessary content push. Taking Taobao as an example, it digs data information about customers' access methods and orders related information in a specific period of time, so as to understand buyers' needs or predict their potential for money, and designs the content and structure of the web page in a targeted way, and assigns each customer a personalized service package or a preferential combination based on appropriate marketing strategies, thus bringing profits to shops on e-commerce platforms [4, 5]. Secondly, data mining can optimize the design of website content. Over the years, e-commerce companies have accumulated a large amount of historical data. How to better utilize this data and explore valuable internal laws and potential customers has become an important means for e-commerce companies to survive and develop.

In order to meet the computational requirements of many data mining systems, technology needs to be adopted in hardware, operating system software, and database systems. These resources greatly increase costs and strain the information technology resources composed of technologists. Not only do distributed and non memory versions of current data mining algorithms need to be developed, but new algorithms also need to be developed, which is a new challenge faced by data mining technology. In order to improve the accuracy and efficiency of e-commerce potential customer data mining, a method for e-commerce potential customer data mining processing based on Apriori association rule algorithm is proposed. This article innovatively adopts a multidimensional tree structure to improve the Apriori association rule algorithm to reduce the frequency of scanning the database; The Vector space model is used to calculate

the similarity between e-commerce potential customers and divide e-commerce potential customers at different levels in customer transaction data. Establish a stickiness evaluation system for potential e-commerce customers as a basic indicator for subsequent mining and processing. Utilize the improved Apriori association rule algorithm to complete data mining and processing of e-commerce potential customers through high-density grids. The study flow module of the method design is shown in Fig. 1:

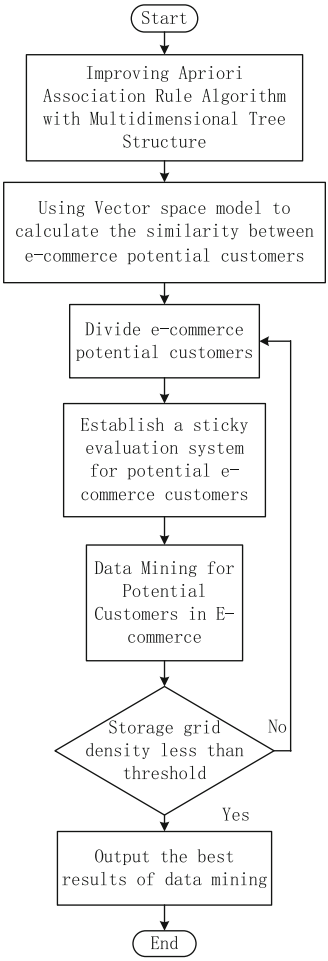

Fig. 1. Research process of each module of the article

2 Improvement of Apriori Association Rule Algorithm

The core idea of the Apriori association rule algorithm is to find all itemsets, ensure that their occurrence times are consistent with the defined minimum support, and then generate strong association rules that can meet the minimum credibility and support.

Then, based on the rules generated from the initial frequency set, generate all rules that only contain items with a set. Once generated, only rules with greater confidence than the given minimum can be left behind. Based on the above characteristics of the Apriori association rule algorithm, it is utilized for e-commerce potential customer data mining and processing [6, 7]. However, during use, this algorithm requires multiple scans of the original database, resulting in a large number of candidate sets being generated, resulting in many redundant rules and low efficiency. This requires further improvements to the algorithm.

The principle of the improved algorithm is to use frequent 1-item sets as candidate item sets, and further expand on this basis by adding a judgment condition Lk. count $\geq$ min sup port to obtain frequent k + 1-item sets. Then, use this as a candidate project set and continuously generate the next frequent project set according to the above steps, ultimately obtaining all project sets. Its advantage is that the frequency of scanning the database decreases, and the number of candidate sets generated continues to decrease. It is more suitable in large-scale databases and can effectively improve the efficiency of extracting frequent itemsets and discovering association rules [8]. To achieve this goal, consider the following principle: assuming there is a data item A, add it to I, and AI, the frequency of the result set (A $\div$ I) is lower than that of I. If I cannot meet the minimum support threshold, then the latter will also not meet, meaning that a set will not pass the test. The following points can be inferred from it.

Improved Apriori association rule algorithm based on multidimensional tree structure. Assuming that an association coefficient is the result of a mapping solution from the Initialization vector to the target vector, multidimensional association rules can be understood as a unified set space composed of multiple association coefficients.

The multidimensional tree structure is used as the basic application structure for large-scale data mining processing, consisting of multiple associated nodes. However, according to the different tasks performed, the corresponding data information objects of each node organization also vary. A Schematic representation of the structure of the association tree organization as shown in Fig. 2.

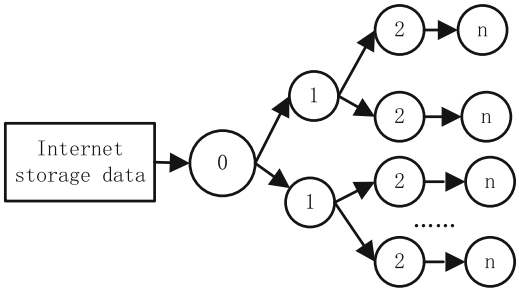

Fig. 2. Schematic representation of the structure of the association tree organization

In Fig. 2, the "0" node serves as the initial structure, responsible for docking with internet stored data, and can directly feed back the information parameters to be mined to the subordinate node structure. As a subordinate structure of node '0', node '1' has certain

data classification capabilities and can be fed back to different storage units according to different encoding forms of data parameters. The "2" node - "n" node serves as the core processing structure of the association tree organization, directly executing data mining instructions, and displaying real-time transmission positions of data information parameters according to the processing results [9]. The actual value of the coefficient "n" varies depending on the length of the association tree organization connection.

Let c denote a randomly selected RFM value definition index, and the inequality condition of the coefficient $c \neq 0$ is constant, and β represent the node definition coefficient in the correlation tree structure. Combining the above physical quantities, the RFM value calculation expression based on the multidimensional association rule can be defined as:

$$z_c = \frac{\beta \cdot x_c}{(\alpha_c + \delta_c)^2 + 1} \tag{1}$$

where x_c represents the scale characteristic value of the e-commerce customer data, α_c and δ_c represent the two unequal multidimensional vector assignment calculation coefficients.

When solving the expression of the RFM value, it is required that the value of the coefficient x_c must belong to the physical interval of $[1, e]$.

In the organization of association trees, the arrangement of feedback nodes not only affects the calculation results of RFM values, but also changes the ability of multidimensional algorithms [10, 11].

Let χ denote the initial assignment of the feedback node distribution coefficient, whose minimum value can only be equal to the natural number "1". ϕ represents the feedback parameter of the data to be mined based on the multidimensional association rules, which is influenced by the solution expression of RFM value. The larger the calculated value of RFM value index is, the larger the actual value of ϕ coefficient is. With the support of the above physical quantities, the joint Formula (1), the multidimensional algorithm expression can be defined as:

$$V = \sum_{\chi=1}^{+\infty} \phi \cdot z_c \cdot \frac{\sqrt{b_1^2 + b_2^2}}{\gamma \cdot \left|\overline{b}\right|} \tag{2}$$

where b_1 and b_2 represent two unequal data information operation features, $\overline{b}$ represents the average value of coefficient b_1 and coefficient b_2, and γ represents the amount of data information extraction parameters based on multidimensional association rules. When constructing the multi-dimensional association rule algorithm, the calculation expression of RFM index should be highly unified. This completes the improvement of the Apriori association rule algorithm.

3 Characteristic Similarity Calculation of E-Commerce Prospect Data

For the behavior data of e-commerce potential customers, it has the characteristics of complex types and large amount of data. In order to accurately and quickly mine the behavior data of target customers, it is necessary to calculate the similarity of the behavior characteristics in advance. The similarity is taken as the scalar value to complete the accurate calculation, and the customers of the e-commerce potential customers at different levels in the customer transaction data are divided. The division structure is shown in Fig. 3:

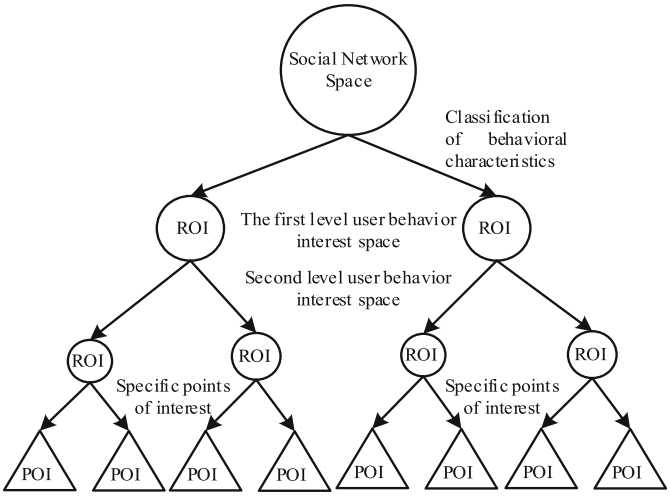

Fig. 3. Schematic diagram of different customer levels

This paper uses the vector space model to calculate the similarity between e-commerce potential customers, each customer is ROI (Region Of Interest region of interest) and POI (Point Of Interest point of interest) vector labeling,$R = [r_1, r_2, \ldots, r_n]$. In order to accurately calculate the increase the number of customer visits, the more similar the behavior expression, a_i is used to indicate the number of the i interest visits. Matrix $V_{m\times n}$ of all customers with the formula:

$$V_{m\times n} = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n-1} & v_{1,n} \\ v_{1,2} & v_{2,2} & \cdots & v_{2,n-1} & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,n-1} & v_{m,n} \end{bmatrix} \tag{3}$$

In the formula, m represents the number of real-time customers; $v_{m,n}$ represents the number of interests of customers in a specific community space; and $v_{m,n-1}$ represents the specific number of interests of customers in the next community space. Take the community network as a space and the customer as a vector value in a dimension in the space. Let the vectors of customers α and β in n dimensions be expressed as U_α and U_β, then the similarity calculation formula between customer α and β is as follows:

$$sim(\alpha, \beta) = \cos\left(U_\alpha, U_\beta\right) = \frac{U_\alpha \cdot U_\beta}{\|U_\alpha\| \|U_\beta\|} \tag{4}$$

Since different customer interest points exist in different spatial neighborhoods, before calculating the similarity of customers under different spatial neighborhoods, it is necessary to cluster the similarity between the two different neighbors:

$$sim_{oreall} = \sum_{i=1}^{H} \mu sim_i \tag{5}$$

In the formula, μ represents the similarity weight, namely:

$$\mu = \frac{\chi_i}{\sum_{i=1}^{H} \chi_i} \tag{6}$$

In the formula, H is the total hierarchy of neighborhood division; χ_i is the weight coefficient. Thus, the similarity of customer data characteristics is calculated.

4 E-Commerce Potential Customer Data Mining and Processing

Based on the similarity results of e-commerce customer data features calculated in the previous section, considering the economic law of e-commerce potential customer stickiness, a e-commerce potential customer stickiness evaluation index system is constructed from aspects: perceived usefulness, perceived ease of use, perceived service, perceived security, perceived interest, and perceived cost performance, as the basic indicators for mining and processing. The evaluation index system of e-commerce customer stickiness is shown in Table 1.

Perceived usefulness represents the customer's belief that using e-commerce platforms is helpful for self-improvement; Perceived ease of use represents the customer's belief that using an e-commerce platform is easy to operate; Perceived service represents that customers believe that using e-commerce platforms has higher service quality [12]; Perceived security represents that customers believe that using e-commerce platforms has a high level of security; Perceived interest represents that customers believe that using an e-commerce platform has a certain level of interest and will increase their interest in using the platform; Perceived cost-effectiveness represents that customers believe that using e-commerce platforms to purchase goods has a higher cost-effectiveness.

Table 1. Evaluation index system of e-commerce customer stickiness

Target layer	The standard layer	Index layer
E-commerce customer stickiness evaluation index	Perceived usefulness	Information channels
		Source of knowledge
		Social help
	Perceived ease of use	Easy to operate
		Retrieve tool efficiency
		The purchase process is simple
		Convenient payment
	Perceived service	Pre-sale seller communication initiative
		Seller delivery timeliness
		Logistics delivery efficiency
		External packaging of goods
		Effectiveness of problem solving
	Perceived safety	Security of website registration and login
		Pay security
		Transaction information security
		Express safety
	Feel fun	Personalized recommendation
		Functional innovation
		The content is rich and interesting

In the improved Apriori association rule algorithm, when dividing the grid of e-commerce customer datasets, the interval segments with consistent boundaries are adjacent interval segments. Make the k-dimensional e-commerce customer dataset $D = \{d_{i1}, d_{i2}, \cdots, d_{iN}\}$, and D have N data points, and divide the i-dimensional e-commerce customer dataset into q intervals, that is, there are $\lceil N/q \rceil$ data points within each interval, that is, equal depth partitioning. Make the data points within each interval consistent. The i-th dimension and jth interval are $H_{ij} = \left(d_{(i)([N/q]*(j-1)+1)}, d_{(i)([N/q]*j)}\right)$, and because the number of data points within each interval is consistent, use the length of the interval to describe the density of the interval, that is $\left|H_{ij}\right| = \left(d_{(i)([N/q]*j)} - d_{(i)([N/q]*(j-1)+1)}\right)$. If the $\left|H_{ij}\right|$ of a certain grid exceeds the corresponding density threshold, then the grid is a high-density grid, and vice versa, it is a low-density grid. If a certain grid and its neighboring grids are both low-density grids, then the grid is treated as noisy data. Let $\left|H_{ij}\right|$ and $\left|H_{i(j+1)}\right|$ in the j + 1 interval be known.

If $|H_{ij}|$ does not obtain a reference source, then the density similarity of adjacent interval segments is $y = \frac{|H_{ij}|}{|H_{i(j+1)}|}$. If $|H_{ij}|$ has obtained a reference source, then $y = \frac{|H_{i(j+1)}|}{|H_{ij}|}$. There are m grids in D memory, and the α The density of each grid is x_α, and the density threshold of the grid is as follows:

$$\varepsilon = \frac{\sum_{\alpha=1}^{m} x_\alpha \lambda}{N} \tag{7}$$

In the formula, the constant is the λ.

The mining result corresponding to the maximum ε value is, and the calculation formula is as follows:

$$k' = \arg\max \left\{ \frac{\sum_{a=1}^{k} \sum_{b=1}^{N} BWP(a, b)}{N} \right\} \tag{8}$$

The sample number of e-commerce customer data is b; the sample number of e-commerce customer data is N; and the mining class number is a.

The principle of improved Apriori correlation rules algorithm is through variable grid segmentation electricity customer data set, comparative analysis of k' and ε, choose more than ε grid, avoid isolated point on the influence of electricity customer viscosity data mining, through the improved Apriori association rules algorithm mining more than ε grid, get the best electricity customer viscosity related data mining results. The input of this algorithm is the data set D of e-commerce customer with N data points, the number of clusters k to expected segmentation, and the similarity threshold v; the output is the data mining result related to stickiness prediction of e-commerce customer. The specific steps of the algorithm to mine the e-commerce customer data set are as follows:

Step 1: Sort each dimension of data in the e-commerce customer dataset using a quick sorting method;

Step 2: Divide each dimension of e-commerce customer data into equal depths;

Step 3: Solve $|H_{ij}|$;

Step 4: Solve the neighboring interval segments of each dimension of e-commerce customer data ρ;

Step 5: Merge adjacent interval segments that meet the conditions;

Step 6: Cycle through the various dimensions of e-commerce customer data in D and merge the eligible interval segments;

Step 7: Calculate the density of the merged mesh;

Step 8: Store the solution record in grid set c;

Step 9: Solve ε;

Step 10: Store grid density exceeding in set d ε Class cluster of;

Step 11: Using the improved Apriori association rule algorithm, divide the high-density grid within d, and output the best results of k e-commerce customer stickiness related data mining.

This algorithm improves the representativeness of the selection of the initial mining center point. It removes the grid density lower than ε in the pruning mode, solves the problem of noise interference, and mines the e-commerce customer data set in the form of grid division, which can handle clusters of different shapes. Thus, the data mining processing of e-commerce prospects based on the Apriori association rules algorithm is completed.

5 Experiments and Analysis

5.1 Experimental Preparation

Taking an e-commerce platform as the experimental object, the public data set of the e-commerce platform contains more than 100,000 data records, which contains three dimensions, respectively, 1 dimension, 2 dimension, 3 dimension, 4 dimension, and 5 dimension, the data set is divided into three subsets according to the data dimension, recorded as subset 1, subset 2, subset 3, subset 4, and subset 5; mining relevant data of the e-commerce platform to improve the economic benefit of the e-commerce platform.

5.2 Comparative Analysis of the Mining Effect

According to the current situation that the data of different dimensions have influence on the mining effect of e-commerce customer behavior data, different dimension data sets are used to mine the user behavior data, the mined information is integrated, and the mining effect is determined through accuracy. Compared with the traditional data mining method of implicit user behavior and the user behavior data mining method based on big data generation, the experimental results are shown in Figs. 4, 5 and 6.

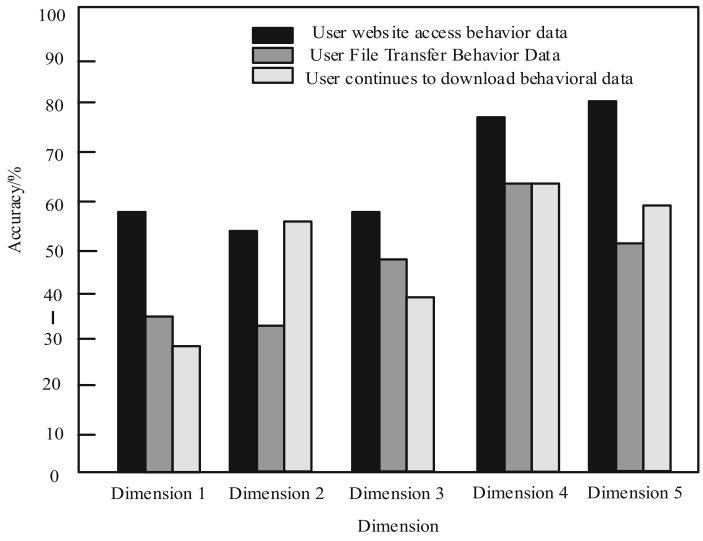

Fig. 4. Mining results of the implicit user behavior mining method

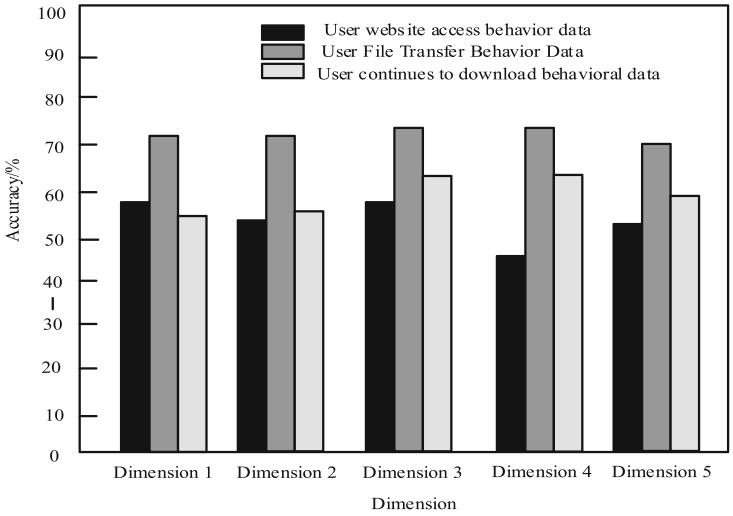

Fig. 5. Mining results of the Big data generation Methods

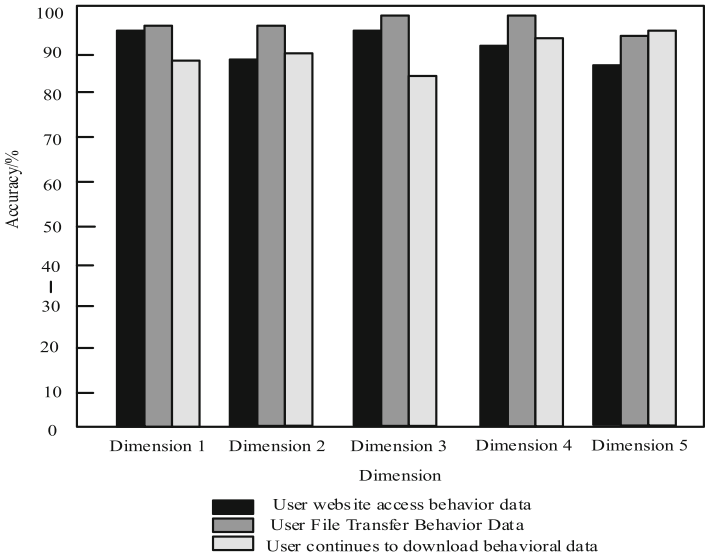

Fig. 6. User behavior mining results used in this paper

As can be seen from Figs. 4, 5 and 6, the user behavior data mined by the method in this paper has the highest accuracy, which is the most consistent with the actual situation. Taking data set 1 as an example, the data mining results are highly consistent with the actual situation, which thus accurately mines the user behavior data during this period; and then observe the other two methods have low accuracy, which is inconsistent with

the actual situation. By comparison, only the method in this paper has a high accuracy of user behavior mining, and the mining accuracy of the other two methods is not high.

6 Conclusion

The data mining method of e-commerce potential customers based on Apriori association rule algorithm is proposed. Improvement of the Apriori association rule algorithm based on the multidimensional tree structure. Calculate the characteristic similarity of the e-commerce potential customer data, and divide the customers among the e-commerce potential customers at different levels in the customer transaction data. From the perspectives of perceived usefulness, perceived ease of use, perceived service, perceived security, perceived interest, the stickiness evaluation index system of e-commerce potential customers is obtained as the basic index of mining and processing. Using the improved Apriori correlation rule algorithm, the data related to e-commerce customer stickiness is mined in the mass e-commerce customer data. Enhance the mining quality of multidimensional data with a high-density grid. The test results prove that this method has high mining accuracy and can be used as the basic data for the e-commerce platform to adjust the marketing strategy. However, due to limitations in research conditions, this method only improves mining accuracy and does not significantly improve mining efficiency. Future research will not only improve mining accuracy, but also improve mining efficiency.

References

1. Zhou, S., He, J., Yang, H., Chen, D., Zhang, R.: Big data-driven abnormal behavior detection in healthcare based on association rules. IEEE Access **8**, 129002–129011 (2020)
2. Menghan, J., Feiteng, L., Zhijian, C., Yu, P.: Robust QRS detection using high-resolution wavelet packet decomposition and time-attention convolutional neural network. IEEE Access **8**(1), 16979–16988 (2020)
3. Qiu, K., Jiang, Y.: A service running data anomaly detection method based on weighted LOF and context judgment in cloud environment. Comput. Eng. Sci. **42**(3), 951–958 (2020)
4. Shuai, L., et al.: Human memory update strategy: a multi-layer template update mechanism for remote visual monitoring. IEEE Trans. Multim. **4**(23), 2188–2198 (2021)
5. Ghosh, D.: Novel trends in resilience assessment of a distribution system using synchrophasor application: a literature review. Int. Trans. Electric. Energy Sys. **31**(5), 129–134 (2020)
6. Jing, L., Siyu, F., Ya-u, Z.: Model predictive control of the fuel cell cathode system based on state quantity estimation. Comput. Simulat. **37**(3), 119–122 (2020)
7. Liu, S., Li, Y., Fu, W.: Human-centered attention-aware networks for action recognition. Int. J. Intell. Syst. **37**(12), 10968–10987 (2022)
8. Du, J., Han, G., Lin, C., Martínez-García, M.: ITrust: an anomaly-resilient trust model based on isolation forest for underwater acoustic sensor networks. IEEE Trans. Mobile Comput. **21**(7), 1684–1696 (2020)
9. Enron, F., Xuren, W., Qiuyun, W., Mengbo, X.: Database anomaly access detection based on principal component analysis and random tree. Comput. Sci. **47**(5), 94–98 (2020)
10. Yanan, S., Xuejing, Z.: An improved outlier detection algorithm and robust estimation. Chin. J. Appl. Prob. Statist. **37**(6), 136–154 (2021)

11. Xianhao, S., Chi, L., Qiong, G., Shaohua, N.: A method for detecting abnormal data of network nodes based on convolutional neural network. Mach. Tool Hydraul. **48**(8), 18–23 (2020)
12. Shiwei, W., Xiaobin, X., Zhongjun, L.: Hierarchical filtering algorithm for distributed abnomaly data based on urban computing. Comput. Integrat. Manuf. Syst. **27**(2), 2525–2531 (2021)