Hindawi
Mobile Information Systems
Volume 2022, Article ID 7202172, 10 pages
https://doi.org/10.1155/2022/7202172

Research Article

Application of Data Mining System in User Network Environment Based on SVM Optimization Algorithm

Yang Yanying

Nanjing Forest Police College, Nanjing, Jiangsu 210023, China

Correspondence should be addressed to Yang Yanying; yhq@lcu.edu.cn

Received 18 August 2022; Revised 13 September 2022; Accepted 21 September 2022; Published 6 October 2022

Academic Editor: Shadi Aljawarneh

Nowadays, different types of data and information combine and interact with each other, forming a complex and huge information network. Using data mining technology, one can effectively obtain the hidden data contained in the data bureau. This technology is the most commonly used way to obtain network target data at present. In this paper, we try to practically apply related algorithms by studying the theory of multi-information fusion. Aiming at the diversity and practicality of the network, the multi-information fusion method was optimized and improved on the basis of the traditional multi-information fusion method. Secondly, a data mining system based on the concept and algorithm of association rules is established, which simplifies the working mode of frequent mining and then improves the data mining model. Finally, an empirical analysis is designed. A group of data samples are selected from the network for preliminary processing, and the data set is brought into the system for testing. From the test results, it can be seen that the algorithm designed in this paper can effectively obtain the target data and works well in a complex network environment, can analyze meaningful data association using user network rules, and provides important guidance for optimizing network information and improving extraction efficiency. This paper combines data mining technology and multi-information fusion technology to conduct in-depth research and further complete the algorithm design by combining the two technologies, which proves the accuracy and processing efficiency of the algorithm.

1. Introduction

With the continuous development of society and technology, new scientific and technological achievements have sprung up like mushrooms after a spring rain. Among them, the most vigorous development is the emerging technology represented by big data and the Internet [1]. The Internet as the carrier of global big data has entered a period of rapid development, and the speed of data growth and the amount of data are also expanding. At present, emerging information technology has been developed by leaps and bounds and has been fully applied by the majority of people [2]. Through the influence of information technology, there are closer interaction channels between major systems. Data and information act as the bridge of system connection, which can fully connect all systems with the network as the carrier, so as to promote the interaction and communication of the system. But in this process, there are also many redundant data points [3]. Therefore, how to effectively deal with useless data in the process of applying information technology has become a key research project. After in-depth research, it was found that the use of data mining technology can effectively deal with redundant data and further convert it into practical data. The reduced data in the application has received widespread attention [4]. The main purpose of data mining is to process and manage multistructured data for large databases. Its scope is quite wide, involving many applications in various fields such as statistical analysis, pattern recognition, social networks, algorithm design, and machine learning. This technology has many applications in machine learning and other fields. It is precisely because of the emergence of data mining technology that the information age has become more computerized, modern society has become more intelligent, and people's lives have become easier. Based on this point, experts from various fields try to combine this technology to create value. Data mining

technology is the most commonly used data processing technology at present, and it is one of the popular application branches based on database technology. It meets the needs of people to acquire useful knowledge and make decisions while working in the context of the current era. Therefore, the use of associated user network rules in a social network can better optimize the process and results of information mining, which is also the goal of this paper [5]. This paper designs and conducts an analysis experiment for user behavior, and the data source is the network data of the target user. Firstly, the original data set is preprocessed, and then the data mining algorithm is applied to complete the processing of the sequence problem. By introducing a multi-information fusion algorithm, the advantages of the algorithm are used to realize the data mining of association rules and the acquisition of user behavior rules, which can provide an information basis for the effective and reasonable application of big data technology [6].

2. Related Work

Fusion is a process of combining, analyzing, and comprehensively processing multisource data or multisensor information to draw new, reliable, and effective conclusions. The applications of multisensor information fusion include multisensor recognition, detection, data association, target tracking, situation assessment, early warning, and so on [7]. The literature shows that more accurate information can be obtained through fusion than from a single input of data, which is also the basic purpose of information fusion technology and the result of synergy. That is, due to the joint action of multiple sensors, the effectiveness of the system has been improved [8]. The literature shows that information fusion is the collection and integration of information from different formats, different sources, and different media with a comprehensive analysis to create more complete, accurate, timely, and effective information, so as to achieve the best consistent estimation of the subject and its attributes [9]. According to the literature, military applications are the source of the birth of multi-sensor data fusion technology, which is mainly used in multistatic radar early warning systems and combat systems, including military target detection, positioning, tracking, and identification (ships), fighters, missiles, etc. The literature shows that not only military but also civil information fusion technology has also made great progress [10]. It is generally used in industrial fields such as robots, smart home manufacturing, transportation, and medical convenience services. It can be used for detection, fault repair, etc. The literature uses a variety of methods to fuse signal change and reconstruction to achieve target signal processing and can realize the feature extraction of the target signal through various domains. The methods used include spectrum analysis, time-domain technology, and so on [11]. A new kind of clustering self-organizing network is designed in the literature, which has many advantages. Compared with traditional networks, it does not need prior information and has a strong anti-interference ability. It can apply relatively simple neurodynamics technology to achieve clustering. Because this is a new technology, it has fewer practical applications, but it has sufficient potential. Data mining is an interdisciplinary technology that includes a large number of other application technologies [12]. It can mine and express hidden data and predict the future trend of data so as to achieve the user's goal by exploring the internal hidden rules of data. The literature shows that data mining algorithms are widely used in many fields, which can promote marketing and industrial production, assist in financial investment, deal with e-government, etc., and they also shine in medicine, biology, and other fields [13]. Such commonly used techniques include envelope anomaly analysis, classification, correlation analysis, clustering, summary regression, time series analysis, etc [14, 15]. The literature discusses the practical problems in the process of sequence analysis. People can sort the original transaction information database, merge the sorted results into the transaction database for operation, and return the results [16]. The literature shows that the average value can be replaced by N adjacent values in the process of trend analysis, followed by simple averaging or weighting, so as to effectively reduce the impact of specific points [17]. If special points are found in the processing process, a smoother and more stable curve can be obtained [18]. This algorithm will obtain uneven curves without obvious specificity [19]. Based on relevant knowledge, the literature divides sequential pattern mining into the following: (1) sorting. In this stage, the transaction subject in the transaction information database is sorted with the primary key and the transaction time as the secondary key, and the sorted data is sorted into the database, and its pattern is transformed into the sorted pattern; (2) big data project processing: at this stage, find all big data sets L, map them to sets of adjacent integers, and each big data set can correspond to an integer; (3) data conversion: at this stage, each entity of the entity sequence value in the database can be replaced with the corresponding big data set object; (4) in the process of sequence processing, large data sets are used to process sequence patterns; (5) maximize the sequence step to obtain the maximum sequence set [20, 21].

3. Theoretical Study of Multi-Information Fusion

3.1. Principle of Multi-Information Fusion. Multi-information fusion is actually an imitation of the process of the human brain processing information, which is similar to other information processing methods. More information about the detected target and environment can be obtained through the information fusion system. At present, there is no unified structure classification form and standard for the information fusion system, but it can be roughly divided into three categories: centralized structure, distributed structure, and hierarchical structure, according to the actual usage habits. All fusion processes are carried out at one fusion center. Each subsystem sends the acquired information and data to the fusion center, and the fusion center conducts a comprehensive analysis and processing of these data and finally obtains the decision-making results. This method has good real-time performance, complete

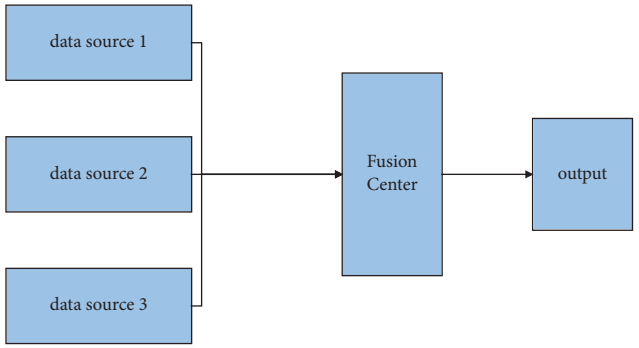

Figure 1: Centralized structure of information fusion.

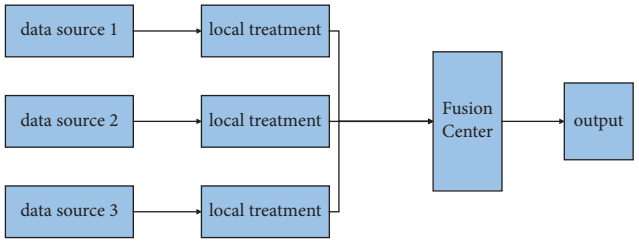

Figure 2: Distributed structure of information fusion.

information processing, and high accuracy. However, it also has certain disadvantages. For example, the processing background of all data is located in the fusion center, which will lead to its complexity. In a certain period of time, due to a large amount of data transmission, it may lead to system collapse and low stability. In this structure, due to the lack of necessary linkages between core sensors, the information and data of each sensor must be sent on time. Therefore, a bus must be provided in the high-speed transmission system and fusion center to achieve a wider range of data transmission. It must also have a more powerful central processing unit to process data, which increases the cost of the project implementation. The structure of the centralized system is shown in Figure 1.

In the distributed structure, the sensor will establish a decision-making system independently and carry out decision-making processing according to the information it receives. After processing, the data will be returned to the fusion center and combined. After the local decision is generated, all decision-making systems jointly estimate and evaluate. This structure does not need the original information, so it greatly reduces the required channel capacity and has a certain fault tolerance. Compared with the actual project, it is easy to realize. This structure has the following disadvantages: the fusion center cannot obtain the target information first. If the decision-making system of a sensor in the system is wrong, the whole system will produce the wrong final results. This means that the uncertainty of integration and the increase of processing projects. Its structure is shown in Figure 2.

The hierarchical structure is to establish a limited connection between each sensor. Some sensors correspond to a fusion node. There can be multiple fusion nodes in the entire system. There are two forms of hierarchy, one without feedback and the other with feedback.

3.2. Overview of Traditional Information Fusion Methods. When using the Bayesian method for multisensor information fusion, the possible decisions of the system must be independent of each other. The Bayesian conditional probability formula is as follows:

$$P\left(\frac{A_i}{B}\right) = \frac{P\left(B/A_i\right)/P\left(A_i\right)}{\sum_{j=1}^{m} P\left(B/A_j\right)/P\left(A_j\right)} \quad i = 1, 2, \ldots, m. \tag{1}$$

If the system has n sensors, the total posterior probability of each decision can be obtained as follows:

$$P\left(\frac{A_i}{B_1 \wedge B_2 \wedge \ldots \wedge B_n}\right) = \frac{\prod_{k=1}^{n} P\left(B_k/A_i\right)P\left(A_i\right)}{\sum_{j}^{m} \prod_{k=1}^{n} P\left(B_k/A_j\right)P\left(A_j\right) \;\; i=1,2,\ldots,m.} \tag{2}$$

Multisensors capture specific features of the target and then simulate and estimate the environment composed of these features to obtain the final synthetic information.

3.3. SVM Optimization Algorithm Design. The SVM training problem can be transformed into a hyperplane optimal classification problem, as shown in the following equation:

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K\left(x_i, x_j\right), \tag{3}$$

$$\sum_{i=1}^{n} y_i a_i = 0, \alpha_i \geq 0, i = 1, 2, \cdots, n.$$

For binary classification problems, the SVM classification discriminant function has the following form:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i^* y_i K\left(x_i, x\right) + b^*\right). \tag{4}$$

Through training, the SVM obtains the corresponding o vector w of the optimal classification hyperplane. This can be represented as a linear combination containing all vectors in the following feature space:

$$w = \sum_{i=1}^{n} a_i y_i \left(x_i\right). \tag{5}$$

We simplify the SVM and replace the initial vector set with a reduced vector set; the equation is as follows:

$$\tilde{w} = \sum_{i=1}^{m} \beta_i \left(x_i\right). \tag{6}$$

The decision function of SVM is shown in formula (7), m is the number of vectors included in the simplified vector set, and $m < n$:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} \beta_i K\left(x_i', x\right) + b^*\right). \tag{7}$$

The processing speed of the classifier can be effectively improved by collecting the simplified vector set with the smallest $m << n$, thereby simplifying the SVM classifier. This process needs to be carried out on the premise of reducing the loss of classification accuracy as much as possible.

There are currently four algorithms:

1.Create a new parsimonious vector and the corresponding weights ($z1$, $\beta 1$) to approximate the following equation:

$$\phi = \sum_{i=1}^{N_x} \alpha_i \varphi(x_i). \tag{8}$$

Then, an iterative construction of the (z_{m+1}, β_{m+1}) approximation vector is performed as follows:

$$\phi_m = \sum_{i=1}^{N_x} \alpha_i y_i \varphi(x_i) - \sum_{i=1}^{m} \beta_i \varphi(z_i). \tag{9}$$

It can be seen from the following equation (10) that this kind of problem can be transformed into a nonlinear multiparameter class optimization problem:

$$\delta = \| \phi_{m-1} - \beta_m \varphi(z_m) \|^2. \tag{10}$$

2. Let the support vector (x_k, y_k) depend linearly on other support vectors in the feature space, namely as follows:

$$k(x, x_k) = \sum_{i=1, i \neq k}^{N_s} c_i k(x, x_i). \tag{11}$$

Then, the corresponding SVM classification discriminant function can be rewritten as follows:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1, i \neq k}^{N_x} \alpha_i y_i k(x, x_i) + \alpha_k y_k \sum_{i=1, i \neq k}^{N_x} c_i k(x, x_i) + b \right). \tag{12}$$

Defining $\alpha_k y_k \gamma_i = \alpha_i y_i \gamma_i$, the classification discriminant function can be rewritten as follows:

$$\begin{aligned} f(x) &= \operatorname{sgn}\left(\sum_{i=1, i \neq k}^{N_x} \alpha_i y_i (1 + \lambda_i) k(x, x_i) + b \right) \\ &= \operatorname{sgn}\left(\sum_{i=1, i \neq k}^{N_x} \overline{\alpha}_i y_i k(x, x_i) \right). \end{aligned} \tag{13}$$

3. A kernel function matrix is given as follows:

$$K = \begin{bmatrix} k_{11} & k_{21} & \dots & k_{l1} \\ k_{12} & k_{22} & \dots & k_{l2} \\ .. & .. & .. & .. \\ k_{1N} & k_{2N} & \dots & k_{\mathrm{lN}} \end{bmatrix} = \begin{bmatrix} \varphi(x_1) \cdot \varphi(s_1) & \varphi(x_2) \cdot \varphi(s_1) & \dots & \varphi(x_l) \cdot \varphi(s_1) \\ \varphi(x_1) \cdot \varphi(s_2) & \varphi(x_2) \cdot \varphi(s_2) & \dots & \varphi(x_l) \cdot \varphi(s_2) \\ .. & .. & .. & .. \\ \varphi(x_1) \cdot \varphi(s_N) & \varphi(x_2) \cdot \varphi(s_N) & \dots & \varphi(x_l) \cdot \varphi(s_N) \end{bmatrix}. \tag{14}$$

Decomposing the kernel function matrix into Km and Kn,

$$\begin{aligned} K_m &= \begin{bmatrix} k_{11} & k_{21} & \dots & k_{l1} \\ k_{12} & k_{22} & \dots & k_{l2} \\ .. & .. & .. & .. \\ k_{1m} & k_{2m} & \dots & k_{lm} \end{bmatrix} = \begin{bmatrix} \varphi(x_1) \cdot \varphi(s_1) & \varphi(x_2) \cdot \varphi(s_1) & \dots & \varphi(x_l) \cdot \varphi(s_1) \\ \varphi(x_1) \cdot \varphi(s_2) & \varphi(x_2) \cdot \varphi(s_2) & \dots & \varphi(x_l) \cdot \varphi(s_2) \\ \dots & \dots & \dots & \dots \\ \varphi(x_1) \cdot \varphi(s_m) & \varphi(x_2) \cdot \varphi(s_m) & \dots & \varphi(x_l) \cdot \varphi(s_m) \end{bmatrix} \\ K_n &= \begin{bmatrix} k_{1m+1} & k_{2m+1} & \dots & k_{lm+1} \\ k_{1m+2} & k_{2m+2} & \dots & k_{lm+2} \\ .. & .. & .. & .. \\ k_{1N-m} & k_{2N-m} & \dots & k_{\mathrm{lN}-m} \end{bmatrix} = \begin{bmatrix} \varphi(x_1) \cdot \varphi(s_{m+1}) & \varphi(x_2) \cdot \varphi(s_{m+1}) & \dots & \varphi(x_l) \cdot \varphi(s_{m+1}) \\ \varphi(x_1) \cdot \varphi(s_{m+2}) & \varphi(x_2) \cdot \varphi(s_{m+2}) & \dots & \varphi(x_l) \cdot \varphi(s_{m+2}) \\ \dots & \dots & \dots & \dots \\ \varphi(x_1) \cdot \varphi(s_{N-m}) & \varphi(x_2) \cdot \varphi(s_{N-m}) & \dots & \varphi(x_l) \cdot \varphi(s_{N-m}) \end{bmatrix}. \end{aligned} \tag{15}$$

The classification discriminant function is given as follow:

$$\begin{aligned} f(x) &= \operatorname{sgn}\left[\sum_{j=1}^{N} \alpha_j y_j k(x, s_j) + b \right] \\ &= \operatorname{sgn}\left[\sum_{j=1}^{m} \alpha_j y_j k(x, s_j) + \sum_{j=m+1}^{N} \alpha_j y_j k(x, s_j) + b \right]. \end{aligned} \tag{16}$$

Make,

$$W^T = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \cdots & \cdots & \cdots & \cdot \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{bmatrix}. \tag{17}$$

Also, with $K_n = W^T K_m$, the discriminant function can be transformed into the following:

$$f(x_t) = \operatorname{sgn}\left(\left(A_m^T + A_n^T W^T\right)K_t + b\right). \tag{18}$$

Only W and m are required. The problem turns into a request,

$$\begin{aligned} &\min\{\varepsilon = \| \Theta' - \Theta \|\} \\ &\text{s.t. } K_n = W^T K_m. \end{aligned} \tag{19}$$

4. Obtain the support vector centroid O_k by kernel clustering.

$$O_k = \sum_{i=1}^{nk} b_{ki}\varphi(x_{ki}) i = 1, 2 \ldots nk = 1, \ldots, \text{cluster num}'. \tag{20}$$

In which,

$$b_{ki} = \frac{\alpha_{ki}}{\sum_{i=1}^{nk} \alpha_{ki} i} = 1, 2 \ldots n, k = 1, \ldots, \text{cluster num}. \tag{21}$$

Calculate the distance from any sample point x_i to the centroid O_k in the feature space.

$$\begin{aligned} \tilde{d}_i^2(O_k, \varphi(x_i)) = K(x_i, x_i) - 2\sum_{p=1}^{nk} b_{kp} K\left(x_{kp}, x_i\right) \\ + \sum_{p,q=1}^{nk} b_{kq} b_{kp} K\left(x_{kp}, x_{kq}\right) k = 1, \ldots, \text{cluster}_{\text{num}}. \end{aligned} \tag{22}$$

For the Gaussian kernel function, there is the following relationship:

$$d_i(z_k, x_i) = -2\sigma^2 \ln\left(1 - \frac{1}{2}\tilde{d}_i^2(\varphi(z_k), \varphi(x_i))\right). \tag{23}$$

An approximation of zk is calculated from this relationship, as a parsimony vector, defined

$$d(\beta_k) = \| \beta_k \varphi(z_k) - \sum_{i=1}^{nk} \alpha_{ki} y_{ki} \varphi(x_{ki}) \|^2, \tag{24}$$

make it the smallest, get the best

$$\beta_k = \frac{\sum_{i=1}^{nk} \alpha_{ki} y_{ki} k(z_k, x_{ki})}{k(z_k, z_k)}, \tag{25}$$

thereby, simplifying is realized.

3.4. Optimal Design of Multi-Information Fusion Method. Generally speaking, the structure of the multisensor information fusion mode is as follows:

Multiple sensors together form a multisensor system, which can provide environmental information and object information about the target. This system is equipped with m fusion nodes, which can fuse the multitarget information. Each fusion node can fuse information according to specific requirements and finally integrate the resulting information Y. In the actual usage process, the choice of fusion method is not fixed and needs to be analyzed in detail. The multi-information fusion node problem based on SVM technology can be expressed as: for an n-dimensional input parameter x, design I independent distributed observation samples are as follows:

$$\begin{aligned} T = \{(x_1, y_1), \cdots (x_l, y_l)\} \in (X \times Y)^l x_i \in X = R^n, \\ y_i \in Y = \{1, 2, \cdots k\}, i = 1, 2, \cdots l. \end{aligned} \tag{26}$$

Find an optimal function f(x) to express the dependency between x and y. In the original space, the resulting decision function is as follows:

$$f(x) = \operatorname{sgn}\left[\sum_{i=1}^{1} a_i^* y_i K(x \cdot x_i) + b^*\right]. \tag{27}$$

To sum up, the multi-information sensor fusion model based on SVM technology can be summarized, and it is found that it has many advantages compared to the traditional multisensor information fusion model: firstly, the model can completely correspond to the general information integration process of output and input nodes. It can convert the nonlinear relationship between the measured parameters and other related parameters, so as to change the fusion form of related information into a mapping relationship and be able to explain part of the content; secondly, the system can convert each parameter into a model function f(x) and transform the problem into a quadratic optimization problem to solve. The value solved in this way can determine the extreme value as the global optimal solution so that the model is consistent; during the test, the accuracy of the system fusion test can be improved by adjusting the kernel function K (x, x) and its related parameters; in addition, this model can effectively solve the nonlinear and high-dimensional problems between related parameters and measurement parameters. In the overall process, since the nonlinear transformation cannot be performed in the input space, it needs to be solved in the feature space. After the solution is obtained, the input of the space vector is compared, so that the nonlinear results are exchanged. In this way, more target work can be done in high-dimensional space, and it can effectively solve the "curse of dimensionality" problem that may occur in other algorithms.

4. Design of Associated User Network Data Mining Algorithm

4.1. Design of Data Mining System. Target knowledge can be obtained by using data mining technology, and data mining can also be carried out in the context of big data. By analyzing the main application process of this technology, we can know that the data mining system can be divided into five main modules, namely preprocessing, data extraction and mining, model evaluation, and output. The structure of the data mining system depends on its organic composition, including the above five functional modules. The architecture diagram of the data mining system is shown in Figure 3.

The key steps of data mining are as follows:

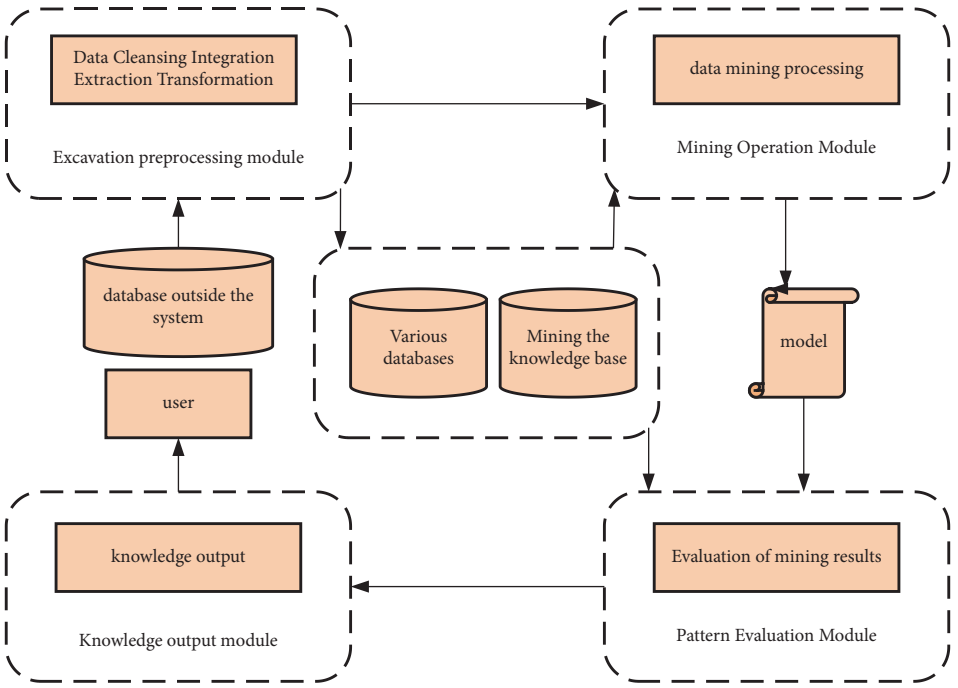

Figure 3: Architecture diagram of a data mining system.

(1) Data preprocessing.

This involves the process of sorting, combining, transforming, and filtering different types of data. Data preprocessing can maintain the database and knowledge base of the system and conduct targeted management. Such databases need to transform, clean, and purify the original data before management and maintenance. The essence of data sorting is data cleaning, which is the basis of data mining preprocessing, including removing redundant data, combining different similar data items, thereby removing data noise, and removing nonidentical data. Data integration is a process of summarizing and combining data, and a database can be used to manage data resources from different sources and databases in real-time to complete the data integration process. The process of data filtering is the act of separating and extracting data from aggregated data and filtering out valuable data that meets the requirements. Data transformation is the secondary integration and data adjustment of the filtered data to integrate the original data elements into a specific form and format more suitable for data mining operations.

(2) Data mining.

The purpose of data mining is to obtain hidden target data and knowledge . This is the most basic content and goal of data-mining technology. This module is mainly composed of a set of fixed functional modules, whose purpose is to perform different types of analysis, including association analysis, characterization, and cluster classification.

(3) Evaluation of the model.

The model evaluation mainly evaluates the results of data mining. Multiple patterns may have been discovered in the previous step, and it is necessary to analyze and compare the user interest points of these standards, and then to evaluate the value of the extracted standards and analyze the reasons for their deficiencies. If the output method is very different from the user's interest, you need to return to the previous process to adjust the parameters and execute it again.

(4) Knowledge output.

In this module, the excavated target data is mainly explained, so as to provide it to the demander in a reasonable and commonly used form. Special data display and interpretation method are used here, which can visualize the result data obtained by mining and provide it to decision-makers for decision-making.

(5) Functional model of data mining system.

Although data mining methods are different, after collecting the application methods of different data mining models, we can develop data mining models by analyzing the needs of users. The development process is based on the summary and analysis of various user problem descriptions and then creates corresponding data models to complete the analysis process and user problem solving process. This process is usually divided into six steps: (1) collect user requirements and complete the definition of various problems; (2) prepare various problems and related basic materials; (3) view and analyze various

data; (4) develop different approaches to solve the problem; (5) validate these generated models to obtain the best model; (6) implement the model.

4.2. The Concept of Association Rules. Based on association rules, the association between item sets in large data sets can be obtained. This process occupies an important position in data mining technology and belongs to an important technical branch.

Unknown associations between unknown relationships and attributes of target objects in the database can be obtained through the study of association rules, and they are hidden in advance, which means that they cannot be obtained through logical database operations (such as table joins) or statistical methods. The inherent attributes of data itself (such as functional dependency) are realized through the symbiotic characteristics between data. Decision making behavior can be assisted by the processing of association rules to facilitate scheme design, website design, business processing, market operation, and decision processing.

In the current era, the use of association rule processing can optimize decision-making, which has attracted the attention of researchers and research institutions in various fields, including big data, artificial intelligence, statistics, and visualization. Through relevant research, relatively fruitful results have been obtained. Because of its easy understanding and simple structure, visualization rules are widely used and become a research hotspot in the field of data mining in recent years, which can effectively analyze the relationship between data.

The confidence level of an association rule is defined as follows:

$$\text{Conf}(\mathbf{R}) = \frac{\text{Sup}(X \cup Y)}{\text{Sup}(X)}. \tag{28}$$

The basic model of association rule mining is shown in Figure 4:

When building the model, the association rule mining algorithm has the following two steps: the first step is to determine all constant datasets in set D according to the minimum support criterion; the second step is to create all association rules based on the constant dataset and minimum confidence threshold.

4.3. Association Rule Algorithm Design. In the process of putting the association rule algorithm into practice, however, since signal level, signal-to-noise ratio, and block error rate are continuous signals, they must be discretized, and the following rules exist for their ternary relationship. The constants set are signal level, signal-to-noise ratio, and block error rate.

As shown in Figure 5, first extract the interference data from the test data, then discretize the data according to the data interval and set the analysis topics of signal level, signal-to-noise ratio and block error rate, realize the analysis algorithm, and obtain the confidence table. Set the appropriate confidence level to obtain the participation relationships of

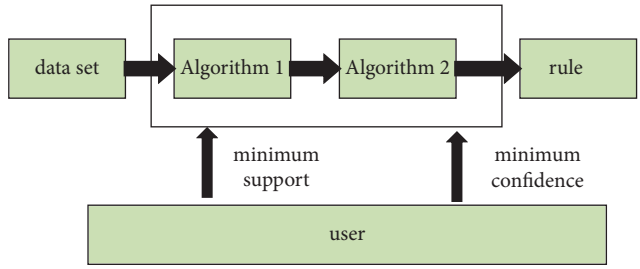

Figure 4: Basic model of association rule mining.

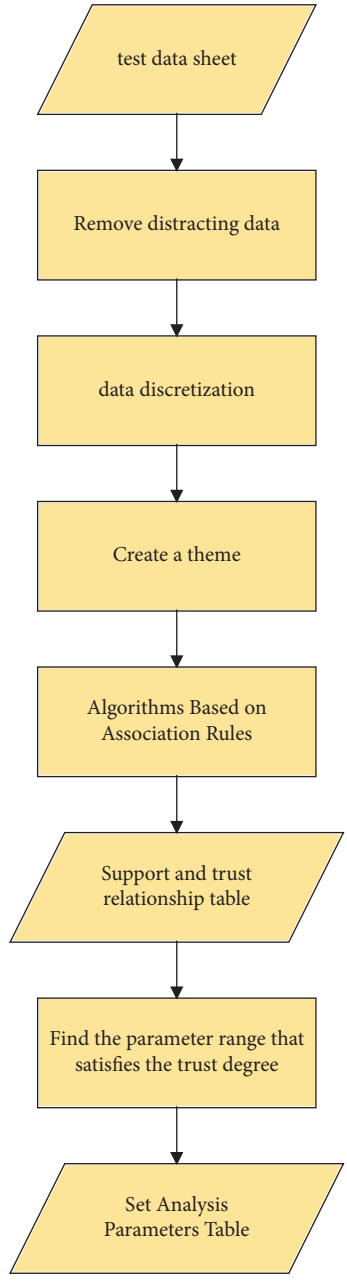

Figure 5: Schematic diagram of association rule module.

signal level, signal-to-noise ratio and block error rate. Relational table values are stored in the database.

4.4. Data Mining Model Construction. Probabilistic knowledge can be used to express the ratio of the support (s) of the association rule X⇒Y to the ratio of tuples with X∪Y in the database, expressed in probability theory as follows:

$$s(X \Rightarrow Y) = \Pr(X \cup Y). \tag{29}$$

The confidence or concentration (α) of an association rule X⇒Y is the ratio of the number of tuples with X∪Y to the number of tuples with X in the database, expressed in probability theory as follows:

$$\alpha(X \Rightarrow Y) = \Pr(Y \mid X). \tag{30}$$

In addition to the two features described above, a third feature can be used to characterize the correlation (σ) between X and Y, which is defined as follows:

$$\sigma_{X,Y} = \frac{\Pr(X \cup Y)}{\Pr(X)P(Y)}. \tag{31}$$

According to probability theory, if $\Pr(X \cup Y) = \Pr(X)\Pr(Y)$ holds, it means that pattern X occurs independently of pattern Y. Therefore, if the correlation is greater than 1, it proves that modes X and Y are positively correlated.

Based on the above three characteristics, the following rules can be defined,

An X⇒Y association rule is said to be valid if the following three conditions are met:

$$\begin{aligned} s(X \Rightarrow Y) &\geq s_{\min}, \\ \alpha(X \Rightarrow Y) &\geq \alpha_{\min}, \\ \sigma_{X,Y} &\geq \sigma_{\min}. \end{aligned} \tag{32}$$

Here, $s_{\min}$, $_{\min}$a>0,$\sigma_{\min}$>1 are all custom data mining thresholds.

5. Implementation and Detection of Associated User Network Data Mining Algorithm

5.1. Sample Processing

5.1.1. Data Cleaning. Because the experimental data is relatively large and these data have some impurities, it may lead to deviations in future data analysis. Some data are not very complete, some even have noise, some have obvious deviations from the real data, and some data is repeated. So, the content of this section is to discuss how to perform data cleaning, smoothing, and denoising. Since the main body of the data is the user, data cleaning is also based on the user. The main principles are as follows:

(1) Delete user data with incomplete information. Such user registration information is incomplete and does not contain explicit personal information.

(2) Remove inactive users who have been inactive for a long time. Although such users may have complete information, they have been silent for a long time and are useless users.

(3) Remove vest users. A vest user is a situation where a user registers multiple IDs on the same website or forum. These vest accounts only serve the user's main account and have no other operations, so these account data will be deleted.

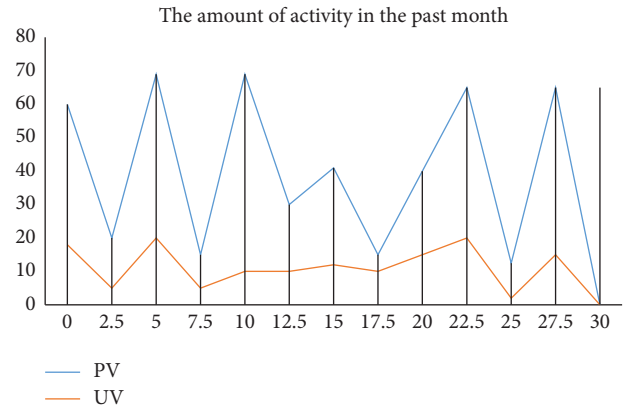

Figure 6: The K-line chart of users' visit data in the past month.

5.1.2. Data Sampling. Although the database size is reduced to 1/5 after the basic database is removed, the database analysis still lacks certain validity and operability because the entire database is still large and there are often hidden data points in the analysis process. Therefore, we use the principles of statistics and use the method of sampling analysis to conduct a sampling analysis of the data.

Advantages of sampling analysis: this is an important statistical research method that shows the relationship and related information of the entire parent group through random sampling of the parent group sample analysis. Its advantage is that it can accurately predict the characteristics of the parent group while reducing the workload, so it takes less time and requires less hardware, greatly improving work efficiency and increasing economic benefits.

Determination of data sampling method and sample size: in the sampling process, three sampling principles must be followed: the principle of validity, the principle of measurement, and the principle of identification and reproducibility; because real comprehensive analysis research must be guided by scientific accuracy. In this study on data sampling, we will choose the principle of stratified sampling. The stratified sampling method divides the population into several layers according to one or more attributes or characteristics, and then draws several samples from each layer and integrates the extracted subsamples into the total sample. Specific operation steps: first, stratify the samples according to specific characteristics or several attributes; second, determine the number of samples that should be drawn from each layer; and finally combine all the obtained samples into the overall sample.

As shown in Figure 6, after statistical analysis, the K-line chart of the user's access data in the last month is obtained. The specific distribution is shown in Figure 6.

Table 1: Test dataset features.

Dataset	Number of transactions	Number of items	Average transaction length
Date 1	3325	65	42
Date 2	8735	116	28

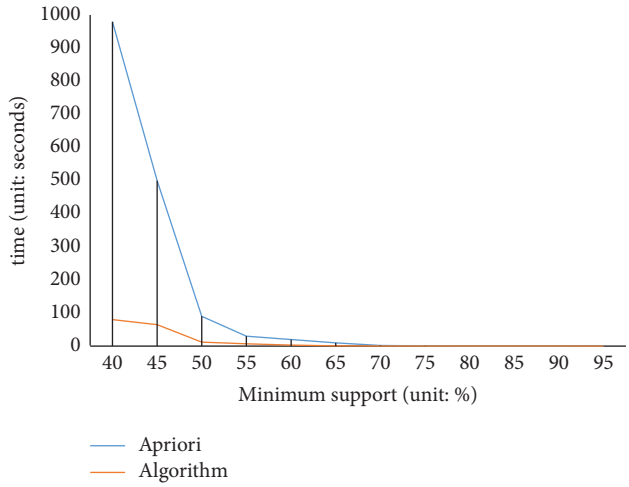

Figure 7: Date1 dataset experiment.

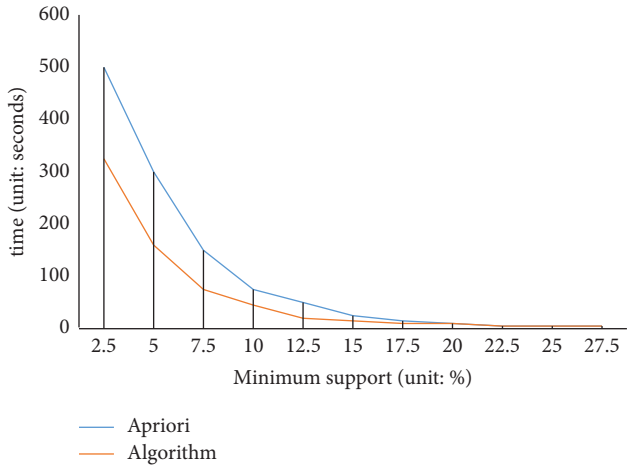

Figure 8: Date2 dataset experiment.

Table 1 lists the eigenvalues of the test dataset.

5.2. Setting up the Experimental Environment. Hardware experimental environment: the hardware requirements of this system are general, and only a few common dual-core machines are needed. Select a computer with better performance as a data mining server to analyze the behavior of network users. Other computers are used by users, and users can generate test access data, access web pages on the network through this data, and store them in the system.

5.3. Experiment Design and Result Analysis. The experimental results of the algorithm running on Date1 and Date2 are shown in Figures 7 and 8.

It can be seen from Figure 8 that the algorithm is better than Apriori in performance, especially in the case of low support.

Experiments show that the algorithm designed in this paper is better than Apriori in terms of time efficiency and has better performance when the minimum support is small.

6. Conclusion

Today, with the rapid development of emerging technologies and society, data mining technology has been fully applied and researched, and it is an emerging branch based on this technology to associate it with multi-information fusion. It contains knowledge in multiple fields and can effectively discover hidden target data and analyze the relationships contained in it. By using the related technologies of extracting network data from associated users, it can better discover valuable information from big data for users, which is a topic worthy of further study. This paper comprehensively analyzes the multi-information fusion algorithm, association rule algorithm, and other related data mining algorithms and then establishes the algorithm as the core algorithm of the user behavior data mining system. Its advantages can better solve related problems. At the same time, because the association rule extraction algorithm is an undirected algorithm, even if the data is preprocessed, some unreasonable rules will be generated. Therefore, effective rules need to be formulated at the same time, so the algorithm still has room for improvement.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

The study was supported by the Jiangsu University philosophy and social science research project "Research on the practice of identifying and tracking epidemic related personnel based on epidemic prediction model"(Grant no.2020SJA0592) and by the Fundamental Research Funds for the Central Universities "Research on named entity depth recognition based on Neural Network"(Grant no. LGYB201703).

References

[1] E. Ahmed, I. Yaqoob, I. A. T. Hashem et al., "The role of big data analytics in Internet of Things," *Computer Networks*, vol. 129, pp. 459–471, 2017.

[2] Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of things and big data analytics for smart and connected communities," *IEEE Access*, vol. 4, pp. 766–773, 2016.

[3] A. Sestino, M. I. Prete, L. Piper, and G. Guido, "Internet of Things and Big Data as enablers for business digitalization strategies," *Technovation*, vol. 98, Article ID 102173, 2020.

[4] C. Zhang, "Classification rule mining algorithm combining intuitionistic fuzzy rough sets and genetic algorithm," *International Journal of Fuzzy Systems*, vol. 22, no. 5, pp. 1694–1715, 2020.

[5] M. Atzmueller, "Data mining on social interaction networks," *Journal of Data Mining & Digital Humanities*, 2014.

[6] X. Zhou and T. Peng, "Application of multi-sensor fuzzy information fusion algorithm in industrial safety monitoring system," *Safety Science*, vol. 122, Article ID 104531, 2020.

[7] S. Chen, L. Huang, J. Bai, H. Jiang, and L. Chang, *Multi-sensor Information Fusion Algorithm with central Level Architecture for Intelligent Vehicle Environmental Perception System*, SAE Technical Paper, Beijing China, 2016.

[8] Z. Zhenxing Li, X. Xianggen Yin, Z. Zhe Zhang, and Z. Zhiqin He, "Wide-area protection fault identification algorithm based on multi-information fusion," *IEEE Transactions on Power Delivery*, vol. 28, no. 3, pp. 1348–1355, 2013.

[9] E. Varol Altay and B. Alatas, "Performance analysis of multi-objective artificial intelligence optimization algorithms in numerical association rule mining," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 8, pp. 3449–3469, 2020.

[10] M. Li, H. Chen, X. Shi, S. Liu, M. Zhang, and S. Lu, "A multi-information fusion "triple variables with iteration" inertia weight PSO algorithm and its application," *Applied Soft Computing*, vol. 84, Article ID 105677, 2019.

[11] M. Zhou, Y. Li, M. J. Tahir, X. Geng, Y. Wang, and W. He, "Integrated statistical test of signal distributions and access point contributions for Wi-Fi indoor localization," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 5, pp. 5057–5070, 2021.

[12] A. Noore, R. Singh, and M. Vatsa, "Robust memory-efficient data level information fusion of multi-modal biometric images," *Information Fusion*, vol. 8, no. 4, pp. 337–346, 2007.

[13] G. Liu and H. Yang, "Self-organizing network for variable clustering," *Annals of Operations Research*, vol. 263, no. 1-2, pp. 119–140, 2018.

[14] A. Bhattacharya, R. T. Goswami, and K. Mukherjee, "A feature selection technique based on rough set and improvised pso algorithm (psors-fs) for permission based detection of android malwares," *International journal of machine learning and cybernetics*, vol. 10, no. 7, pp. 1893–1907, 2019.

[15] T. L. Lai, "Sequential analysis: some classical problems and new challenges," *Statistica Sinica*, vol. 7, pp. 303–351, 2001.

[16] O. Lindwall, G. Lymer, and C. Greiffenhagen, "The sequential analysis of instruction," *The Handbook of Classroom Discourse and Interaction*, vol. 23, pp. 142–157, 2015.

[17] E. V. Altay and B. Alatas, "Chaos numbers based a new representation scheme for evolutionary computation: applications in evolutionary association rule mining," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 5, Article ID e6744, 2022.

[18] R. Bakeman and J. R. Brownlee, "The strategic use of parallel play: a sequential analysis," *Child Development*, vol. 51, no. 3, pp. 873–878, 1980.

[19] E. V. Altay and B. Alatas, "Differential evolution and sine cosine algorithm based novel hybrid multi-objective approaches for numerical association rule mining," *Information Sciences*, vol. 554, pp. 198–221, 2021.

[20] A. Farouk, A. Alahmadi, S. Ghose, and A. Mashatan, "Blockchain platform for industrial healthcare: vision and future opportunities," *Computer Communications*, vol. 154, pp. 223–235, 2020.

[21] B. S. Maitner, B. Boyle, N. Casler, R. Condit, J. Donoghue, and S. M. Durán, "The bien r package: a tool to access the Botanical Information and Ecology Network (BIEN) database," *Methods in Ecology and Evolution*, vol. 9, no. 2, pp. 373–379, 2018.