Eng Int Syst (2024) 3: 165–174

Engineering
Intelligent Systems

Method for Constructing Network Intrusion Detection Model Based on Improved Apriori Algorithm

Linlin Wu*

Department of Information Engineering, Yellow River Conservancy Technical Institute, Kaifeng 475003, China

While the efficient information age brings convenience to people, it is also accompanied by myriad network dangers. In order to detect and respond to frequent network attacks, this research introduces the improved Apriori algorithm and K-means algorithm, and establishes a network intrusion detection model based on these two algorithms. The PR curves of the two algorithms before and after the improvement indicate that the AP value of the improved Apriori algorithm is 0.9972, which is significantly higher than that before the improvement, 0.9324. In addition, two datasets, testSet and Iris, were used to test the three improved K-means algorithms. Results show that the L-kmeans algorithm has the highest clustering accuracy, with an improvement of about 19% on the testSet dataset, and the accuracy of the L-kmeans algorithm on the Iris dataset is increased by about 14%. Finally, the performance of the improved model is verified by the detection efficiency of Snort. The most significant change in this improvement method is that in terms of false detections, the number of dangerous behaviors identified as normal data and the number of normal data behaviors identified as dangerous behaviors are significantly reduced by 53.0% and 32.0%, respectively. At the same time, the number of undetected dangerous behaviors and normal data behaviors also decreased by 37.4% and 36.5%, respectively. The accuracy, stability and efficiency of the model are verified by simulation experiments.

Keywords: network security; intrusion detection, association rules; data mining; Apriori

1. INTRODUCTION

With the advent of the era of big data and the rapid development of network information technology, computer networks have now developed into a mainstream productivity tool. However, with the complexity of network technology, more and more network security problems have begun to emerge [1]. Due to the large amount of data, a high degree of data sharing must occur in order to use this data efficiently. However, during the sharing process, many criminals, high-risk viruses, etc. will invade the network with the shared information, resulting in confusion, loss of data, and even paralysis of the entire system [2]. Although several network security detection systems are available at present, due to the increasing volume and updating of network traffic data, traditional security detection methods cannot adapt to the network attack behavior of such update speed. As a widely used security protection technology, intrusion detection technology has been applied in various network security protection. However, most of the current intrusion detection technologies are relatively old technologies, and there have been many loopholes in network security systems, and a lot of human and financial resources have been wasted [3]. Therefore, it is necessary to explore a faster, more accurate and more efficient method of monitoring network security. The Apriori algorithm is a type of association rule algorithm, and one of the earliest of its kind [4]. This algorithm uses the method of layer-by-layer search and iteration to mine data, and then determines the association rules between the data according to the information between the mined data. According to previous research, the Apriori algorithm can quickly find potential attack signals in massive

*Email of corresponding author: wl830221yrcti@163.com

network information and respond quickly [5]. Therefore, this research will establish a network intrusion detection model based on the Apriori association rule algorithm, and conduct simulation experiments to study the effect of this method when applied to network intrusion detection.

2. RELATED WORKS

The Apriori algorithm is the most widely-used association rule data mining algorithm, and has been applied in many fields. Sun et al. [6] proposed an Apriori association rule algorithm for dynamic early warning of the original target course grades of college students, and improved it to mine a large amount of data from the system and finally obtain management rules. The results show that the system ultimately promotes the improvement of college students' academic performance and achieves the purpose of talent training. When the Iorliam et al. [7] conducted a Nigerian terrorism forensic investigation, they proposed to use the Apriori algorithm in their study. Findings show that most attacks in Nigeria are successful and that most Nigerian terrorists do not die by suicide. Kusak et al. [8] introduced Apriori association rules and the K-means clustering algorithm for cluster analysis in order to evaluate the landslides in the Karahel River and to map them geologically. The analysis results show that the average horizontal sliding of the area affected by the landslide is 25.56m and the average precision, precision and recall are 0.78, 0.73 and 0.68 respectively. Zhang et al. [9] proposed the Apriori association rule algorithm when studying the difference in disease patterns caused by urban-rural differences among the elderly in China, and performed logistic regression processing and cross-sectional analysis on the data. The experiment achieved good results, successfully analyzed the difference in disease patterns caused by urban and rural factors, and verified the feasibility of this method. In order to study the impact of financial turmoil on the cryptocurrency market in 2020, José et al. [10] introduced the Apriori association rule algorithm to find the association rules between different currencies, in order to determine whether the price or the volume of the currency constitutes a rule. The results of the study show that it is found that before the price of cryptocurrencies falls, the association rules are generally formed by these prices, and then, the transaction volume dominates to form the association rules.

In order to explain the defects and dependencies in yacht production in detail, Tacjana et al. [11] proposed to use the Apriori association rule algorithm to model and analyze them. Simulation experiments show the presence of many dependencies, which are not obvious to technicians, but occur with a high level of probability. The proposed research results may lead to improvements in the planning process for production tasks. When studying network security issues, Fu et al. [12] proposed a deep learning network intrusion detection model, integrating an attention mechanism and a bidirectional long-short-term memory network for network intrusion detection to analyze it. Simulation experiments show that the model's accuracy and F1 score are better than those of other comparison methods, reaching 90.73% and 89.65%, respectively. Alavizadeh et al. [13] proposed a deep Q-learning (DQL) reinforcement learning method and established a Q-learning model when they studied the problem of network intrusion detection. Experimental results show that the proposed DQL is very effective in detecting different intrusion classes and outperforms other similar machine learning methods. Abed et al. [14] proposed a machine-learning model that combines autoencoders with a class of support vector machines when researching network intrusion detection methods. They found that testing on NSL-KDD and KDD99 datasets produces greater accuracy, with total accuracies of 96.24% and 99.45%, respectively. When Nureni et al. [15] studied the activities of illegal network intrusions to the network, they proposed a solution to implement intrusion detection on the Hadoop MapReduce framework by using the improved hash-based Apriori algorithm. Experimental results show that this method is a reliable and effective means of detecting network intrusions. Previous studies have shown that today's data mining methods still have a lot of problems in terms of network security detection, so this study introduces a more accurate and efficient Apriori algorithm to model and analyze it, in order to create a more secure network.

3. NETWORK INTRUSION DETECTION MODEL BASED ON IMPROVED ASSOCIATION RULE ALGORITHM

3.1 Apriori Algorithm and its Improvement on Additional Constraints

An association rules algorithm was first proposed and designed to discover the association rules between different commodities. These rules show the purchasing habits of consumers, and merchants can offer more specifically-targeted products and merchandising according to consumers' purchasing habits, so as to achieve destocking and so on. Therefore, improving the accuracy, efficiency, and applicability of association algorithms has always been a hot issue in academic research [16]. The Apriori algorithm is a type of association rules algorithm that has been widely respected by academic circles since its introduction in the late 20th century. The algorithm uses the method of layer-by-layer search and iteration to mine data, and then determines the association rules between the data according to the information between the mined data.

Association rules $A \Rightarrow B$ are expressed by A which is called the association antecedent, and B which is the association consequent. It is assumed that all sets in the project are shown in Formula (1).

$$I = \{I_1, I_2, I_3, \ldots, I_k\} \tag{1}$$

If $\exists X \subset I$, $Y \subset I$, it is called X, Y is the item set. If the item count in it is k, then it is called k item set; if the set consisting of all items in the target library is, $I = \{I_1, I_2, I_3, \ldots, I_k\}$, $D = \{T_1, T_2, T_3, \ldots, T_k\}$ is the transaction library, where the transaction is shown in Formula (2).

$$T_i = \{I_{i1}, I_{i2}, I_{i3}, \ldots, I_{ik}\} \tag{2}$$

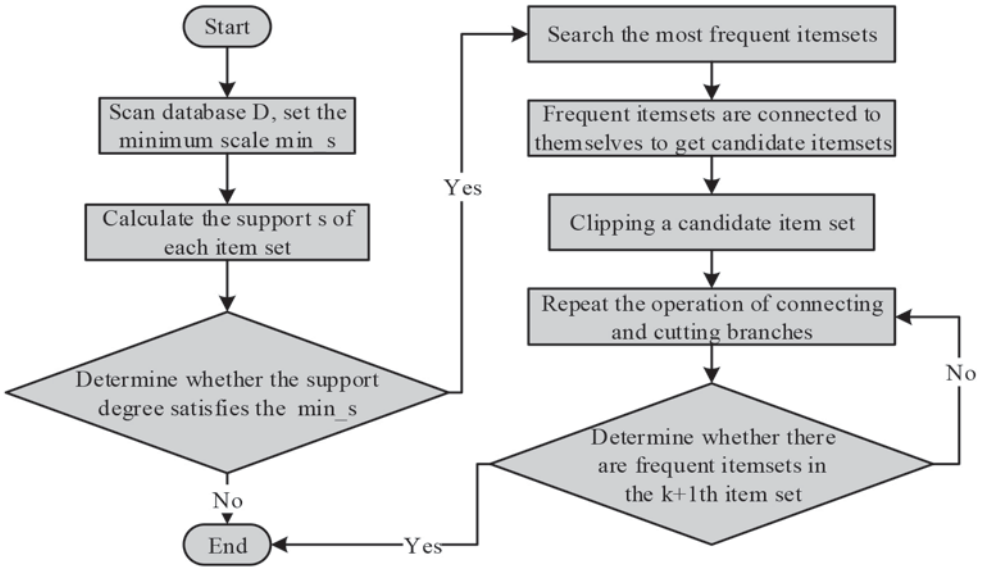

Figure 1 Flowchart of Apriori algorithm implementation.

In Formula (2), $\forall I_{ij}(j \in [1,k]) \subseteq I$. Define $A \Rightarrow B$ to denote the association rule, where A, B are shown in Formula (3).

$$\{A, B | A \subset I,\ B \subset I,\ A \cap B = \Phi\} \tag{3}$$

Assuming that A and B are both included by the transaction of $s\%$ in the target library, then the support of $A \Rightarrow B$ is said to be $s\%$, as shown in Formula (4).

$$Support(A \Rightarrow B) = P(A \cup B) = s\% \tag{4}$$

Assuming that there are $c\%$ transactions in the target library containing the itemset A that also contain the itemset B, then $c\%$ is said to denote the confidence level of $A \Rightarrow B$ as shown in Formula (5).

$$Confidence(A \Rightarrow B) = P(B|A) = \frac{Support(A \cup B)}{Support(A)} = c\% \tag{5}$$

If $\exists X \Rightarrow Y$ the support s and confidence c are both greater than the set minimum support min_s and minimum confidence min_c, it is called a strong association rule. However, when $X \Rightarrow Y$, the Apriori algorithm faces two main problems. First, the method also generates a large number of candidate sets in addition to the target set, which is wasteful; second, when calculating association rules, the algorithm needs to scan the target library repeatedly. Although these two defects have been effectively overcome after years of research, to date, there is another defect that has not been effectively addressed. *Snort* is a tool used to monitor in real time the traffic that goes in and out of a network. However, there is no effective method for improving the algorithm for intrusion detection [17]. If the original Apriori algorithm is used directly without improvement, currently, since the algorithm mines only high-frequency itemsets, many dangerous attack behaviors and security information will be detected at the same time, and the detection system will produce many false positives. The sensitivity of the monitoring system is greatly reduced, and because the first group of targeted items do not have any association rules linking them, any information found through this method is of little or no use for future decision-making. If only the minimum support degree, $\min_s$, is adjusted, this cannot effectively improve this situation, so constraints should be added to increase the degree of association. The specific Apriori algorithm implementation flowchart is shown in Figure 1.

As shown in Figure 1, the minimum support is first assumed to be min_s and all transaction datasets are traversed $D = \{T_1, T_2, T_3, \ldots, T_k\}$. Then, according to the support calculation formula, the support of all itemsets is calculated s, and the frequent itemset with the smallest support is found. The obtained frequent itemsets are then self-connected to obtain candidate sets. Then the obtained candidate set is pruned to obtain a new frequent itemset. The self-connection and pruning operations are repeated on the newly-obtained frequent itemsets until $k+1$ and no frequent items appear in the th item set. Then the operation is terminated and the kth frequent itemset is obtained [18].

3.2 K-means Data Mining Improved Algorithm

The K-means clustering algorithm (k-means) has a very high sensitivity to cluster centers, and the diversity of clustering results is also extremely rich. In this study, L-K-means algorithm is selected as an improved algorithm for data mining. In this study k, the non-parametric probability density algorithm combined with the traditional K-means was used to calculate the density of the points. The points with a density greater than 15% were screened out, and then these points were used to preliminarily determine the coordinates of the cluster centers. In terms of k quantity, this study adds the cluster radius as a new parameter. Each time a cluster center is determined, a circular area is delineated with the added cluster radius until all the circular areas are large enough to cover the entire dataset. When the cluster centers are

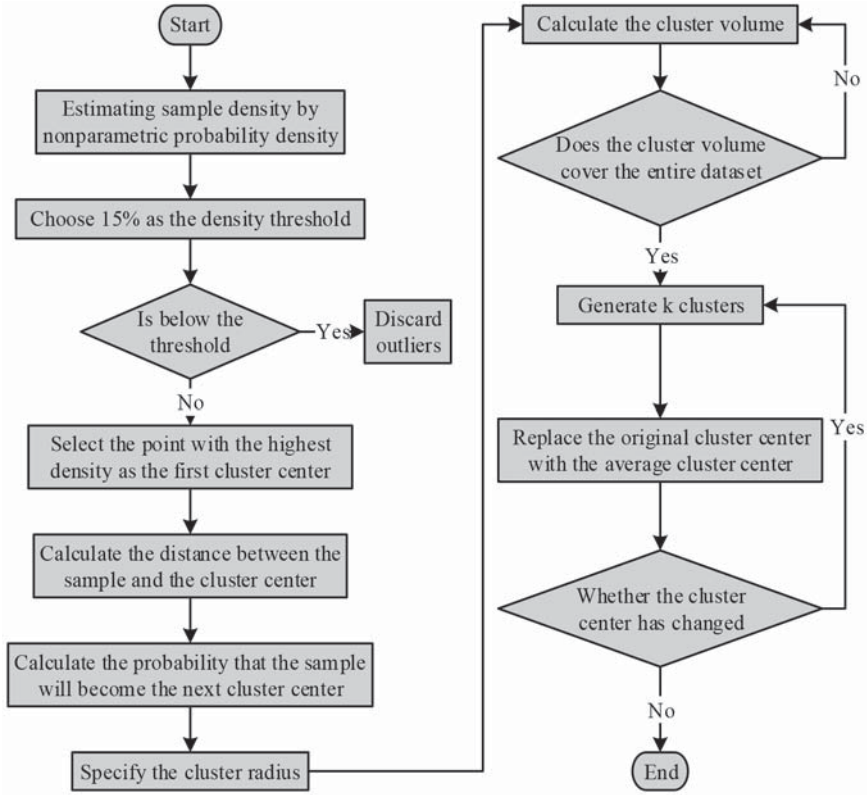

Figure 2 Specific flow of L-K-means algorithm improvement.

determined by k, the location and size of clusters are more precise and accurate, and the clustering results are also more stable. The improvement of this clustering algorithm has three main characteristics. First, the use of the non-parametric probability density algorithm used to determine the center of the point is highly convincing; second, the fixed threshold is more accurate; and third, by increasing the clustering radius to delimit the circular region, the optimised algorithm has more constraints in the specified region compared to the traditional k-means, resulting in more stable results, and the method also provides a good basis for subsequent research [19]. The specific process is shown in Figure 2.

As shown in Figure 2, when improving the L-K-means algorithm, firstly, the sample density of each point on the spatial dataset is calculated using the nonparametric probability density function. As shown in Formula (6).

$$P_n(X_i) = \frac{K_n/n}{V_n} \tag{6}$$

In Formula (6), $P_n(X_i)$ denotes the probability density of each point X_i in the dataset, K_n denotes the number of points in the space of volume V_n, and V_n denotes the volume of the space containing K_n. The probability density of each point in the dataset is the probability density of each point in the dataset. And it must be satisfied that $P_n(X_i)$ converges to $P(X_i)$, as shown in Formula (7).

$$\lim_{n\to\infty} K_n = \infty \tag{7}$$

$$\lim_{n\to\infty} \frac{K_n}{n} = 0 \tag{8}$$

As shown in Formulas (7) and (8), K_n, n the change directions of the two remain the same, and n the change rate is much larger than K_n that. When calculating X_i the distance to all points in the dataset, X_i the previous K_n close sample point is as shown in Formula (9).

$$D_i = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2},\ (j = 1, 2, \ldots, n,\ j \neq i) \tag{9}$$

Then according to X_i the average distance of the neighboring points R as V_n the radius, as shown in Formula (10),

$$R = d_i(X_i, X_t) = \frac{1}{t}\sum_{j=1}^{t} \|x_i - x_j\| \tag{10}$$

In Formula (10), it d_i represents the neighbor distance and X_t represents the set of adjacent t points. After the sample density is calculated, points with a density below 15% are discarded. At the same time, the first cluster center is calculated, as shown in Formulas (11)–(13).

$$R \geq R_{0.1n} \tag{11}$$

$$c_1 \in V_{R_{\min}} \tag{12}$$

$$c_1 = \frac{\sum_{i\in c_1} X_i}{N(c_1)} \tag{13}$$

In Formula (11), $R_{0.1n}$ denotes the threshold value of the radius. In Formula (12), $V_{R_{\min}}$ denotes the region with the smallest radius. In Formula (13), $N(c_1)$ indicates the number of data in the region with radius c_1.

Then, when finding the next cluster center, first calculate the distance from the sample point to the first cluster center, as shown in Formula (14).

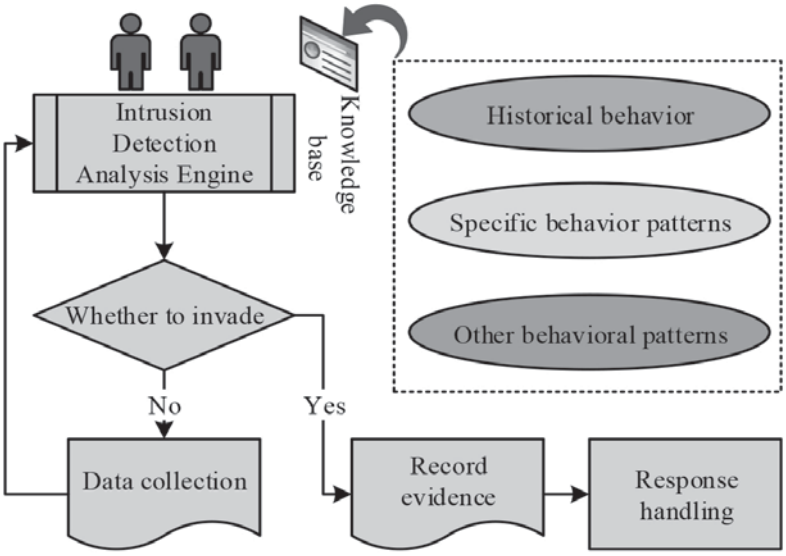

Figure 3 Framework of basic intrusion detection system.

$$D_i(x) = \sqrt{(x_i - x_1)^2 + (y_i - y_1)^2},\ (i = 1, 2, \ldots, n) \quad (14)$$

After calculating the distance, then calculate the probability that each sample point is selected as the next cluster center, as shown in Formula (15).

$$Q_i = \frac{D_i(x)^2}{\sum\limits_{x \in X} D_i(x)^2} \quad (15)$$

Then select the second cluster center by the roulette method, and find all the cluster centers according to the above steps. Finally, clusters are formed, and then the cluster centers are continuously updated, and the above operations are repeated until the cluster centers are stable [20].

3.3 Establishment of Network Intrusion Detection Model

Given the huge information databases of networks, an effective network intrusion detection technology must be used to detect in real time any anomalies in network traffic to ensure the security of the system. Usually, the function of network intrusion detection technology is to identify the legitimacy of network behavior. Hence, in the process of research, we can reasonably describe the behavior of the system in terms of characteristic patterns, and accurately judge the legitimacy and safety of a behavior. The traditional network intrusion detection structure model is shown in Figure 3.

As shown in Figure 3, the intrusion monitoring system comprises three main parts: data extraction, intrusion analysis, and corresponding processing. The functions of each part are different. The data extraction part carries out the crucial first task of the system: the collection of data information prior to the follow-up work that involves, importantly, intrusion analysis, based on the classification and sorting of various types of data. Feature information it is extracted from this data to determine whether there is an abnormality. The third part of the system is the response function. When the system detects a danger, this part starts to process the potentially dangerous information [21].

The network intrusion detection model established in this research is based on the improved Apriori algorithm and K-means algorithm. First, data mining is carried out on the log of the website server, and then the Apriori algorithm and the K-means algorithm are applied to analyze the browsing behavior of all IP addresses, in order to find potential intrusion behaviors and block them in time. The structure of the proposed network intrusion detection model based on the improved Apriori algorithm and K-means algorithm is shown in Figure 4.

As shown in Figure 4, this model consists of five parts: data collection, data preprocessing, data analysis, detection agent and detection response. This model has four main functions: firstly, sorting and classifying the IP addresses in the website server log; secondly, sorting out the abnormal state in the log record; thirdly, the abnormal state obtained is analyzed by association rules; fourthly, the K-means algorithm is used to perform cluster analysis on the classified data and, finally, the response function is used to respond.

For data collection, network data tools are usually used to obtain data directly from websites to generate datasets that can be used for intrusion detection [22]. The collected data is then pre-processed, because the directly collected data is usually messy and incomplete, and cannot be used immediately. In order to ensure the accuracy of detection, the data must be preprocessed [23]. When performing preprocessing operations, it is usually necessary to read data to perform operations such as cleaning, classification, and transformation, so that the data can meet the requirements of subsequent operations. Data cleaning is usually done to ensure the integrity and smoothness of the data, so that it can have a unified format and there is no duplicate data. Data classification involves combining data with the same attribute(s) into the same dataset. Data conversion is done to convert data with different formats, standards and specifications into a form that enables the data to be mined.

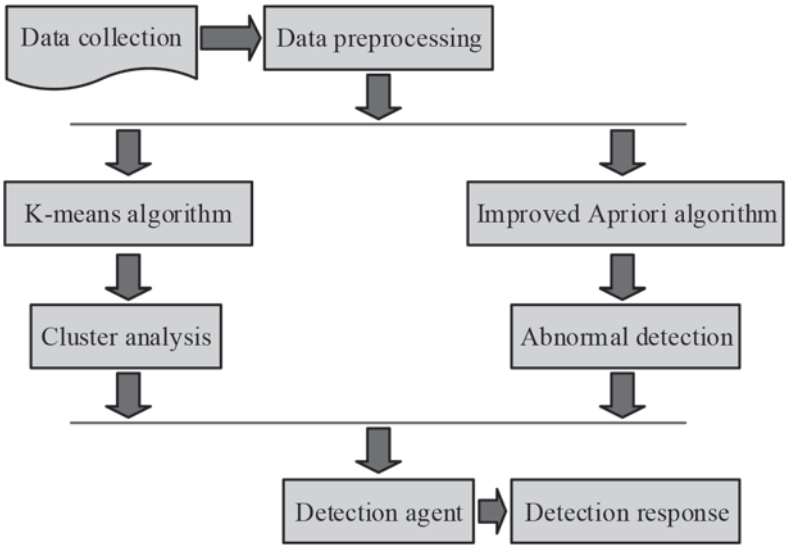

Figure 4 Network intrusion detection model based on improved Apriori algorithm and K-means algorithm.

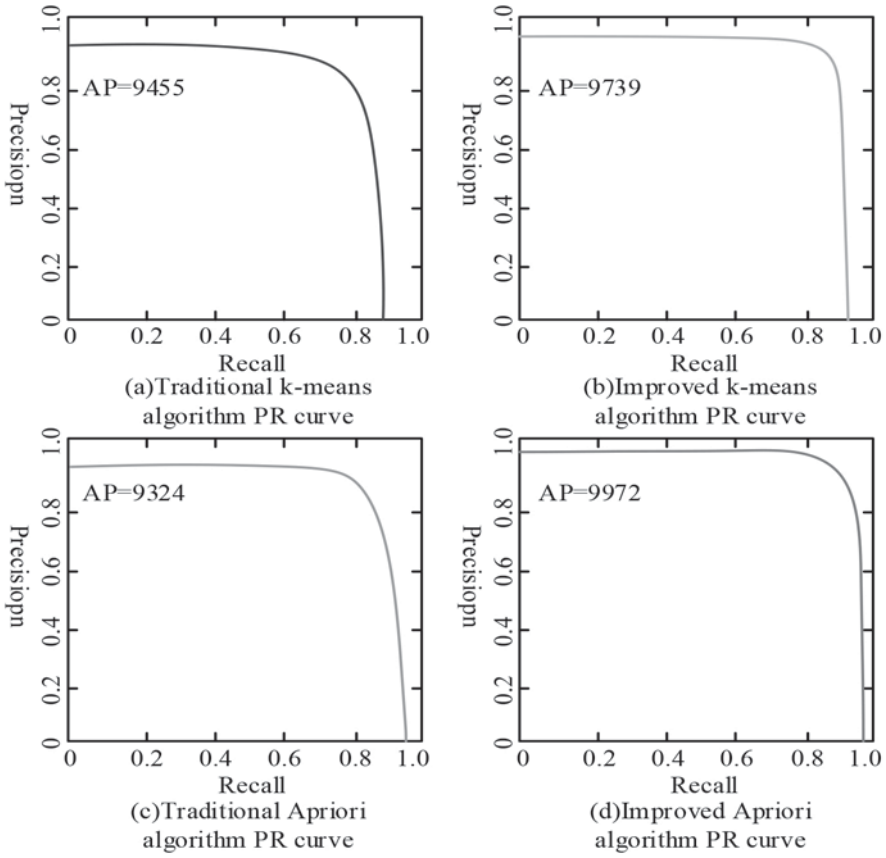

Figure 5 PR curve of the improved Apriori algorithm and K-means algorithm.

4. SIMULATION EXPERIMENT OF NETWORK INTRUSION DETECTION MODEL

In experiments, the threshold for dividing the positive and negative cases of the predicted value is usually set to 0.5, and this must be done in advance. When the predicted value is greater than or equal to 0.5, the corresponding sample at this time is classified as a positive example; similarly, when the predicted value is less than 0.5, the corresponding sample is classified as a negative example. If the set threshold value is continuously changed, the PRecision and Recall values also change continuously, and the PR curve reflects the different test results. Figure 5 shows the PR curve of the test results.

Figure 5(a) shows the PR curve formed by the test results of the traditional K-means algorithm, Figure 5(b) shows the PR curve formed by the test results of the improved K-means algorithm, and Figure 5(c) is the traditional. The PR curve

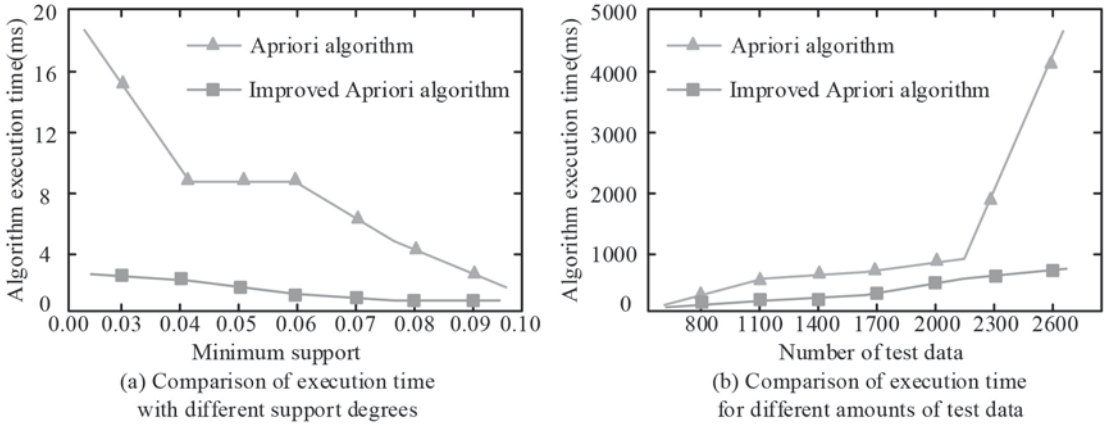

Figure 6 Plot of the results obtained with the improved Apriori algorithm.

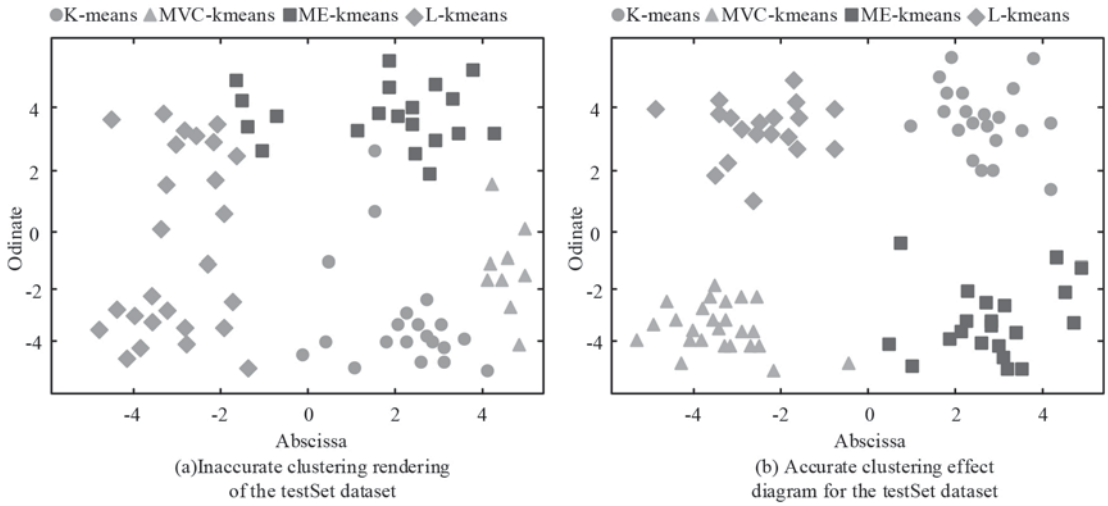

Figure 7 Test results for testSet dataset.

formed by the test results of the Apriori algorithm, Figure 5(d) is the PR curve formed by the test results of the improved Apriori algorithm. It can be seen from Figure 5(a) and 5(b) that the improved K-means algorithm is obviously better than the previous one in terms of accuracy and detection speed, and the AP value of the improved K-means algorithm is 0.9739, which is significantly higher. It is 0.9455 before improvement, and has better clustering effect. It can be seen from Figure 5(c) and 5(d) that the improved Apriori algorithm has a better balance between precision and recall, and is more stable in comparison. And the AP value of the improved Apriori algorithm is 0.9972, which is significantly higher than the 0.9324 before the improvement, and the detection accuracy and efficiency are greatly improved.

After the PR curve analysis, the improved Apriori algorithm is analyzed. In the experiment, different minimum support degrees and different data amounts were changed to observe the calculation results. The results are shown in Figure 6.

Figure 6(a) shows the running time of the two algorithms on different minimum support degrees, and Figure 6(b) shows the running time of the two algorithms when tested on different amounts of data. When the fixed amount of data is constant and the minimum support is small, the execution time of the two algorithms is very different. As the minimum support increases, the running time of both algorithms decreases significantly. In general, the running time of the traditional Apriori algorithm is longer than that of the improved Apriori algorithm; when the fixed minimum support invariant data volume is 500 to 1000, the running time of the two algorithms is not much different. As the amount of data continues to increase, the running time of the traditional Apriori algorithm increases significantly more than that of the improved Apriori algorithm. To sum up, the operating efficiency of the improved Apriori algorithm has been greatly improved compared with that before the improvement.

The LK-means algorithm was tested on the testSet dataset and the Iris dataset. Figure 7 shows is a set of simulation experiment results performed by the LK-means algorithm on the testSet data set.

Figure 7(a) shows the inaccurate clustering effect for the testSet dataset, and Figure 7(b) shows the accurate clustering effect for the testSet dataset. It can be seen from the figure that the clustering accuracy of the four algorithms is relatively high. The clustering accuracy of the unimproved K-means algorithm is above 74%, and the other three improved algorithms exhibit some improvement. The MVC-k-means algorithm is improved by about 8%, the ME-k-means algorithm is improved by about 13%, and the

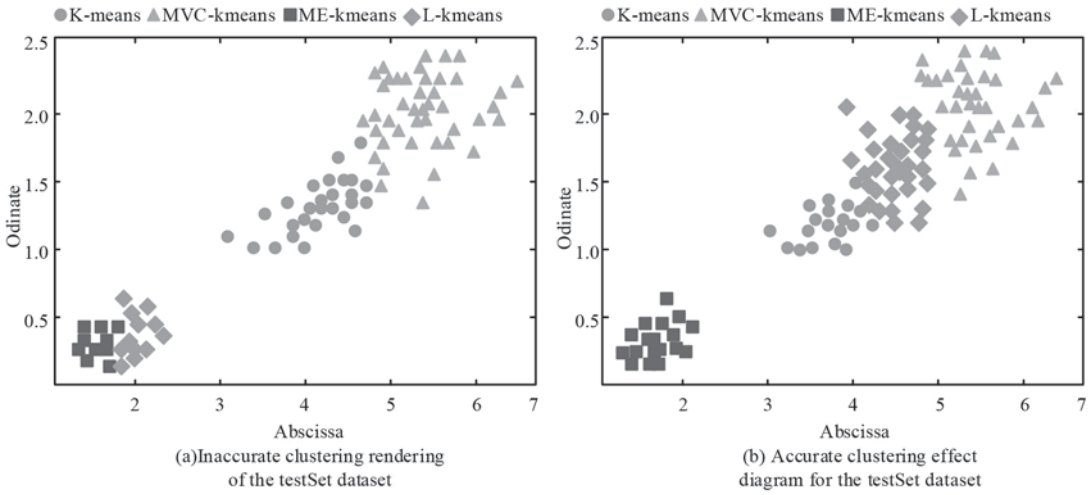

Figure 8 Test results for Iris dataset.

Table 1 Comparison of Snort time spent before and after improvements.

Test data (10^4)	System detection time (s)		Time saving (s)	Testing time saving rate (%)
	Detection method			
	After improvement	Traditional		
0.1	13	13	0	0
0.5	24	24	0	0
1	89	92	3	3.26
1.5	337	349	12	3.44
2	4699	4987	288	5.78
2.5	6011	6321	310	4.90
3	8357	8743	386	4.41
4	11694	12119	425	3.51

L-k-means algorithm is improved by about 19%. As evident. the L-k-means algorithm has the most obvious improvement in accuracy on the testSet data set. Figure 8 is a set of simulation experiment results performed by the LK-means algorithm onon the Iris data set.

Figure 8(a) shows the inaccurate clustering effect of the Iris dataset, and Figure 8(b) is the accurate clustering effect of the Iris dataset. It can be seen from the figures that the clustering accuracy of the four algorithms is relatively high. The clustering accuracy of the unimproved K-means algorithm is about 58%, and the other three improved algorithms have also achieved better accuracy. Among them, the MVC-k-means algorithm and the ME-k-means algorithm have similar improvement effects, both improved by about 4%, and the L-k-means algorithm has improved by about 14%. When comparing the three improved algorithms, it is found that the L-k-means algorithm still has the most obvious improvement in accuracy when applied to the Iris data set.

In this study, the Apriori algorithm was improved and combined with the improved K-means algorithm to increase the detection efficiency of Snort. In this simulation experiment, the detection efficiency of Snort before and after the improvement was also compared and analyzed. The results are shown in Table 1.

As shown in Table 1, when the test data is less than 5000, the improved detection method has basically no effect on the detection efficiency of Snort. As the amount of detection data increases, the level of detection efficiency changes. When the detection data reaches 10,000, the detection efficiency of Snort is improved by 3.26% compared with the traditional detection method. When the detection data reaches 20,000, the detection time of Snort in the improved detection method is reduced by 288s, and the detection efficiency is increased by 5.78%. When the number of detections gradually increased to 40,000, the improved detection time reached 11694s. Compared with the traditional detection method, the saving rate was 425s, and the detection efficiency was increased by 3.51%. It can be seen from the results that this improved method significantly improves the detection efficiency of Snort.

In terms of detection efficiency, it is found that the efficiency of the improved detection method has been significantly improved, and the change in the detection accuracy of the improved network intrusion detection method will be further analyzed. Due to the huge amount of data, the position detection gap method will be used to characterize the detection accuracy. The specific detection results are shown in Table 2.

As shown in Table 2, the content of this test mainly includes normal data and dangerous behaviors that are missed, detected dangerous behaviors and normal data and dangerous behaviors, and the number of dangerous behaviors that are wrongly identified as normal data. From the data in the table,

Table 2 Comparison of the detection accuracy of Snort before and after the improved intrusion detection method.

Network intrusion detection methods			
Test content	After improvement	Traditional	Alignment detection gap (%)
Dangerous behavior detected	233791	222270	5.18
Normal data detected	51072	49455	3.27
Risky behavior is identified as normal data	4288	9120	−53.0
Normal data is identified as risky behavior	420	618	−32
Risky behaviors of missed inspections	13347	21334	−37.4
Missing normal data	3299	5198	−36.5

it can be seen that the intrusion detection methods optimized by the improved Apriori algorithm and the K-means algorithm have increased the number of normal data and dangerous behavior detections, with an increase rate of 3.27% and 5.18%, respectively. In addition, the number of undetected dangerous behaviors and normal data decreased by 37.4% and 36.5%, respectively. The most significant improvement is that in terms of false detections, there is a significant reduction in the number of risky behaviors identified as normal data and normal data identified as dangerous behaviors, by 53.0% and 32.0%, respectively, thereby greatly improving the accuracy of network intrusion detection.

5. CONCLUSION

In order to construct a more complete network intrusion detection system, this research introduces the Apriori algorithm and K-means algorithm. A network intrusion detection model is established based on the two algorithms, and simulation experiments are carried out. It can be seen from the PR curve that the improved Apriori algorithm and the K-means algorithm achieve a better balance between the precision rate and the recall rate, and are more stable in comparison. And the AP value of the improved Apriori algorithm is 0.9972, which is significantly higher than that of 0.9324 before the improvement, and the AP value of the improved K-means algorithm is 0.9739, which is significantly higher than the 0.9455 before the improvement, which has great detection accuracy and efficiency. promote. By changing different minimum support degrees and different amounts of data, it can be seen that the running time of the traditional Apriori algorithm is longer than that of the improved Apriori algorithm; the three improved K-means algorithms are tested through the testSet data set and the Iris data set. It can be seen that, of the four algorithms, the L-k-means algorithm has the highest clustering accuracy. The accuracy of the L-k-means algorithm improved by about 19% onwhen applied to the testSet dataset, and the accuracy of the L-k-means algorithm onapplied to the Iris dataset increased by about 14%. Finally, the performance of the improved model is verified by the detection efficiency of Snort. When the number of detections was gradually increased to 40,000, the improved detection time reached 11694s. Compared with the traditional detection method, there was a saving of 425s, and the detection efficiency was increased by 3.51%. Also, it has greatly improved in terms of false detection, missed detection, and the number of detections. To sum up, the network security detection model based on the improved Apriori algorithm and the K-means algorithm is a great improvement on the previous model.

Data Availability Statement

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

Conflict of Interest

The authors declare that they have no competing interests.

Funding Statement

There is no funding for the study reported in this paper.

REFERENCES

1. Chen M, Yin ZX. (2022). Classification of cardiotocography based on the Apriori algorithm and multi-model ensemble classifier. *Frontiers in Cell and Developmental Biology*, 10:888859–888859.
2. Greco C, Fortino G, Crispo B, Choo KKR. (2023). AI-enabled IoT penetration testing: state-of-the-art and research challenges. *Enterprise Information Systems, 17*(9), 2130014.
3. Bai H. (2022). Key factor mining method of distribution network equipment operational efficiency based on Apriori and CNN. *Energy Reports, 8*(S3):533–538.
4. Zeng J, Jia B. (2022). Live multiattribute data mining and penalty decision-making in basketball games based on the Apriori algorithm. *Applied Bionics and Biomechanics, 2022*:789–791.
5. Ma H, Ding JJ, Liu M, Liu Y. (2022). Connections between various disorders: combination pattern mining using a priori algorithm based on diagnosis information from electronic medical records. *BioMed Research International, 2022*:317–321.
6. Sun GH, Guo SC, Hao G, Yang WB. (2021). Dynamic early warning system of College Students' target course performance based on improved Apriori algorithm. *Journal of Computational Methods in Sciences and Engineering, 21*(6):1779–1795.

7. Iorliam A, Dugeri RU, Akumba BO, Otor S. (2021). A forensic investigation of terrorism in Nigeria: An Apriori algorithm approach. *Journal of Information Security, 12*(04):270–280.
8. Kusak L, Unel FB, Alptekin A, Celik MO, Yakar M. (2021). Apriori association rule and K-means clustering algorithms for interpretation of pre-event landslide areas and landslide inventory mapping. *Open Geosciences, 13*(1):1226–1244.
9. Zhang CC, Xiao SJ, Shi L, Xue YQ, Zheng X, Dong F, Zhang JC, Xue BL, Lin H, Ouyang P. (2021). Urban–Rural differences in patterns and associated factors of multimorbidity among older adults in China: A cross-sectional study based on a priori algorithm and multinomial logistic regression. *Frontiers in Public Health, 9*:61–62.
10. José BHC, Andrés GM, Miguel APV. (2021). Study of the behavior of cryptocurrencies in turbulent times using association rules. *Mathematics, 9*(14):1620–1620.
11. Tacjana NR, Michal L, Paweł S. (2020). Application of a priori algorithm in the lamination process in yacht production. *Polish Maritime Research, 27*(3):59–70.
12. Fu YF, Du YS, Cao ZJ, Li Q, Xiang W. (2022). A deep learning model for network intrusion detection with imbalanced data. *Electronics, 11*(6):898–898.
13. Alavizadeh H, Alavizadeh H, JangJaccard J. (2022). Deep Q-learning based reinforcement learning approach for network intrusion detection. *Computers, 11*(3):41–41.
14. Abed S, Alshayeji HM, AlSulaimi M, Jaffal R. (2022). Network intrusion detection with auto-encoder and one-class support vector machine. *International Journal of Information Security and Privacy (IJISP), 16*(1):1–18.
15. Nureni AA, Tolulope JA, Sanjay M, Rytis M, Robertas D. (2019). Network intrusion detection with a hashing based a priori algorithm using hadoop mapreduce. *Computers, 8*(4): 86–86.
16. Xie HY. (2021). Research and case analysis of a priori algorithm based on mining frequent item-sets. *Open Journal of Social Sciences, 09*(04):458–468.
17. Cheng KC, Huang MJ, Fu CK, Wang KH, Wang HM, Lin LH. (2021). Establishing a multiple-criteria decision-making model for stock investment decisions using data mining techniques. *Sustainability, 13*(6):3100–3100.
18. Yang T. (2021). Erratum: characteristics predicting a high caregiver burden in patients with vascular cognitive impairment: using the a priori algorithm to delineate the caring scenario [corrigendum]. *Risk Management and Healthcare Policy, 14*:3195–3195.
19. Wardani DW. (2021). Measuring positive and negative association of Apriori algorithm with cosine correlation analysis. *Baghdad Science Journal, 18*(3):554–564.
20. Jhang KM, Wang WF, Chang HF, Chang MC, Wu HH. (2021). Characteristics predicting a high caregiver burden in patients with vascular cognitive impairment: using the Apriori algorithm to delineate the caring scenario. *Risk Management and Healthcare Policy, 14*:1335–1351.
21. Vashisht A, Holy C, Shah S, Elangovanraaj N., Johnston S., Coplan P. (2020). PIN108 Using the Apriori Algorithm to Identify Risk Factors Associated with Survival and Mortality Among COVID-19 Patients. *Value in Health, 23*(S2): S561-S562.
22. Gaurav A, Gupta BB, Panigrahi PK. (2023). A comprehensive survey on machine learning approaches for malware detection in IoT-based enterprise information system. *Enterprise Information Systems*, 17(3), 2023764.
23. Mathur ON, Li CC, Gonen B, Lee KJ. (2022). Application of representation learning-based chronological modeling for network intrusion detection. *International Journal of Information Security and Privacy (IJISP), 16*(1):1–32.