Research Article

Yongkuan Zhu, Gurjot Singh Gaba, Fahad M. Almansour, Roobaea Alroobaea, and Mehedi Masud*

Application of data mining technology in detecting network intrusion and security maintenance

https://doi.org/10.1515/jisys-2020-0146
received December 30, 2020; accepted March 09, 2021

Abstract: In order to correct the deficiencies of intrusion detection technology, the entire computer and network security system are needed to be more perfect. This work proposes an improved k-means algorithm and an improved Apriori algorithm applied in data mining technology to detect network intrusion and security maintenance. The classical KDDCUP99 dataset has been utilized in this work for performing the experimentation with the improved algorithms. The algorithm's detection rate and false alarm rate are compared with the experimental data before the improvement. The outcomes of proposed algorithms are analyzed in terms of various simulation parameters like average time, false alarm rate, absolute error as well as accuracy value. The results show that the improved algorithm advances the detection efficiency and accuracy using the designed detection model. The improved and tested detection model is then applied to a new intrusion detection system. After intrusion detection experiments, the experimental results show that the proposed system improves detection accuracy and reduces the false alarm rate. A significant improvement of 90.57% can be seen in detecting new attack type intrusion detection using the proposed algorithm.

Keywords: data mining, intrusion detection, k-means improved algorithm, security maintenance

1 Introduction

With the development of the Internet, the Internet has become an important part of human work and life. Especially after the realization of globalization and informatization, many enterprises, government agencies, and individuals carry out various businesses and operations on the open Internet, such as Financial companies carry out online banking, etc. [1]. People use the Internet to facilitate work and life, and at the same time, change people's work lifestyle and meet the needs of modern people for being fast, efficiency, and convenience. People use the Internet to obtain the information they need, as we all know The Internet is open and shared. These two characteristics make the resources integrated into the Internet rich, but at the

* **Corresponding author: Mehedi Masud,** Department of Computer Science, College of Computers and Information Technology, Taif University, Taif, KSA, e-mail: mmasud@tu.edu.sa
Yongkuan Zhu: Department of Information Engineering, Henan Polytechnic, Zhengzhou, Henan, 450046, China, e-mail: yongkuanzhu1@gmail.com
Gurjot Singh Gaba: School of Electronics and Electrical Engineering, Lovely Professional University, Phagwara 144411, India, e-mail: gurjot.17023@lpu.co.in
Fahad M. Almansour: Depertment of Computer Science, College of Sciences and Arts in Rass, Qassim University, Buraydah 51452, Saudi Arabia, e-mail: f.almansour@qu.edu.sa
Roobaea Alroobaea: Department of Computer Science, College of Computers and Information Technology, Taif University, Taif, KSA, e-mail: r.robai@tu.edu.sa

same time it also brings hidden dangers to the security of the Internet itself [2]. Intrusion Detection System (IDS) is one of the technologies to improve network security. It collects information on certain key points in the network for technical analysis to find abnormalities or intrusion attacks. Dynamic security technology is for network detection, alarm, and response. The main function of an IDS is to detect intrusive behaviors and identify intrusive events. In essence, it is to classify network behavior data [3]. The network data are divided into normal and abnormal data, and the judgment of abnormal behavior is intrusion detection. The IDS can monitor the hosts, running applications, and their status in the system in a comprehensive and real-time manner. It can also detect intrusion behaviors of internal and external attacks in the system in real time, actively identify intrusions, issue an alarm, etc. The above characteristics make it play an important role in network security. Figure 1 illustrates the architecture of a basic intrusion detection system.

Data mining technology is useful in finding the patterns for a large dataset using various multidisciplinary techniques like machine learning (ML), statistics, and artificial intelligence. Data mining is basically about discovering the hidden and unpredictable relationships among the data by the detection of data patterns, knowledge extraction, and revealing the unknown information. The insight of these data mining strategies can be used to evaluate the probability of future events which can be used in various fields of marketing, scientific discovery, fraud and intrusion detection, etc. The usage of data mining techniques for intrusion detection is significantly gaining importance in the current scenario for predictive analysis of suspicious intrusions. The data mining-based IDSs are effective in identifying the data of user's interest and can better predict the outcomes for future security maintenance [4,5]. Intrusion detection was initially introduced in the year 1980 by Anderson [6] and later, the concept of data mining advanced the detection process. This advancement involves a great deal of consideration in the IT sector for the discovery of knowledge in the society [7,8]. Data mining can easily access the useful information from the large volume of data [9]. In Figure 1, the intrusion model has been centrally placed in order to capture all the data packets and transmit them over a network. The data collected from the database is also preprocessed in order to make it free from the redundancy and noise [10]. These preprocessed data are then further analyzed and categorized according to the severity measures, and based on the data state, the alarm is raised to the security maintenance administrator to handle the situation. The literature related to IDS provides various alternatives for addressing this problem; however, the security maintained is still challenging due to complex data classification and data labeling tasks.

The shortcoming of intrusion detection technology has been addressed in this work, to facilitate the performance of computer and network security systems. This work proposes an improved k-means algorithm and an improved Apriori algorithm which are applied in the data mining technology for the detection of network intrusion and security maintenance. The classical dataset has been exploited in this work by

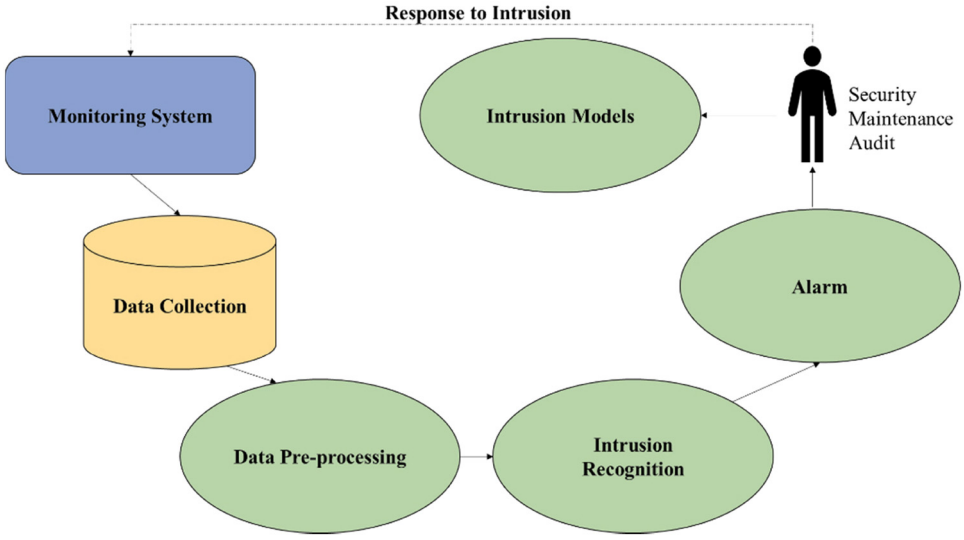

Figure 1: Architecture of a basic intrusion detection system.

performing the experimentation with the improved algorithms. The contribution of improved algorithms has been tested experimentally in terms of detection rate and false alarm rate. Significant improvement has been noticed in the outcomes obtained and the improved algorithm advances the detection efficiency and accuracy using the designed detection model. When the improved and tested detection model is then applied to a new IDS, a significant improvement is seen in the detection accuracy while simultaneous reduction in the false alarm rate.

The remaining paper is organized as follows: Section 2 presents the literature review of the state of the art work in the field of intrusion detection. The dataset description and data preprocessing are detailed in Section 3 followed by the experimental results and analysis in Section 4. Section 5 presents the conclusion of the entire research work followed by the future research scope.

2 Literature review

Deng and others analyzed the characteristics of network security and security issues and discussed the system framework of Internet security and some key security technologies, including key management, authentication and access control, routing security, privacy protection, intrusion detection, and Fault tolerance and this article introduces the current problems of the Internet of Things in network security and points out the necessity of intrusion detection. Several intrusion detection technologies are discussed, and their applications in the Internet of Things architecture are analyzed. We compared the applications of different intrusion detection technologies and looked forward to the next phase of research. Using data mining and ML methods to study network intrusion technology has become a hot issue. It is difficult to improve the detection rate of network intrusion detection by a single type of feature or detection model. The performance of the proposed model is verified by a public database [11]. Zhang J., with IDS as the research object, established an IDS model based on data mining, obtained experimental results, and reached relevant experimental conclusions. At the same time, it was compared with traditional IDS, and six experiments were carried out. The output results of the detection rate, false negative rate, and false positive rate of two different methods in six experiments were obtained. The experiment draws a conclusion: the IDS using data mining has better network protection and security performance, and the detection ability of network vulnerability intrusion is stronger. This research provides a new way for the detection and research of network protection security loopholes [12]. Bagaa M. et al. proposed a novel ML-based security framework that can automatically deal with extended security aspects related to the IoT domain. The framework utilizes software-defined networking and network function virtualization enablers to mitigate different threats. The AI framework combines monitoring agents and ML-based AI-based reaction agents. These ML-Models are divided into network pattern analysis and anomaly-based intrusion detection in IoT systems. The framework uses supervised learning, distributed data mining systems, and neural networks to achieve its goals. The experimental results prove the effectiveness of the scheme. In particular, the use of data mining methods to distribute attacks is very successful in detecting attacks with high performance and low cost. Regarding an anomaly-based IDS for the Internet of Things, the experiment has been evaluated in a real smart building scenario using a type of SVM. The detection accuracy of anomalies reaches 99.71%. A feasibility study was conducted to determine the current potential solutions to be adopted and to promote the development of the study in the direction of open challenges [13].

Ektefa et al. [14] evaluated the IDSs using the machine leaning platforms, and among C4.5 and Support Vector Machine (SVM) methods, C4.5 performs better. A hybrid Particle Swarm Optimization (PSO) was proposed by Holden and Freitas [15] to deal with the nominal attributes for intrusion detection. Ardjani et al. [16] applied the combination of PSO and SVM to optimize the ML algorithm performance and 10-fold cross validation is done for accuracy estimation. Multidimensional dataset was dealt with to remove the redundancy and inconsistency in the feature vectors and provide better classification outcomes [17]. Two-class classification method was proposed by Wael et al. [18] to classify the normal situation from the attack class. This method was effective in providing the least error and better detection rate comparative to its counterparts. New form of attack patterns was captured by Petrussenko [19] for the detection of attacks in

the network packets and it was found that higher detection rate was achieved by utilizing a preprocessor module in between [20,21]. Improving IDS by ML needs large sets of labeled data for attack scenarios and normal scenarios. Acquiring such sets is costing time and effort. Also, it can be done by experts. A large set of pre-labelled training data was utilized by an unsupervised method to produce less accuracy; therefore, the authors dealt with the semi-supervised algorithm to provide better accuracy detection rates. A fuzzy clustering-based approach was proposed by the authors in ref. [22], while evaluating both the Euclidean distance and statistical cluster properties to identify the possibility of intrusion. Wang and Megalooikonomou [23] presented a co-training framework to improve the intrusion detection in case of unlabeled data. This method provides the lower error rate, while incorporating an active learning technique to improve the system performance. Higher detection rates were achieved by utilization of a semi-supervised learning mechanism which yields comparatively reduced false alarm rate [24]. In order to improve the accuracy rate of intrusion detection, a tri-training SVM algorithm has been used [25]. A tree-based clustering technique was proposed by Li et al. [26], which finds the clusters among the intrusion detection dataset. This method works efficiently faster comparative to other algorithms. A combination of misuse and anomaly detection was utilized by the authors in ref. [27] for determining the cost function of the semi-supervised strategy involving SVM classifiers. For addressing the distributed denial of service attacks, various advents have been offered [28]. Fu and Papatriantafilou [29] proposed a perspective proactive algorithm to divide the network into set of clusters and the intrusion is detected using these clusters. Fu et al. [30] presented a synchronous communication-based method for intrusion detection and reducing the false alarming rates.

The usage of data mining combined with the ML methods has been made for intrusion technology in various related works. All the proposed models in the literature try to fit into the prescribed dataset and reduce the administrator workload by improving the detection of attacks. However, from this literature survey, it was found that it is challenging to improve the detection rate of network intrusion detection by a single type of feature or detection model, as used in the current existing methods. These methods are still needed to be improved and there are manifolds possibilities of improvement in IDSs.

3 Introduction to KDD cup99 data set

KDDcup99 is a data set specially used to evaluate the situation of research IDSs. It is a data set collected by Lincoln Laboratory during intrusion detection and evaluation of DARPA networks in 1998. It contains various types of data containing different intrusion categories. The data set contains a total of 4.9 million data; here, only 10% of the data are taken from “kddcup.data_10. percent” [31]; that is, 396,473 normal behavior data and abnormal behavior data. Since all the data in the KDDcup99 data set are the original network data packets obtained by Tcpdump, the data packets are converted into ASCII format.

Intrusion detection evaluation KDDcup99 data set mainly has the following types of attacks:

(1) Denial Service (Denial Service);
(2) Illegal remote host access R2L attack (Remote to Local);
(3) Local unprivileged users conduct unauthorized access U2R attacks (User to Root) of local super users;
(4) Reconnaissance and detection of probing attacks;
(5) Data transmission data attack.

The data provide 41 characteristics, divided into three different categories: traffic characteristics, content characteristics, and other characteristics.

3.1 Basic features

The format conversion of each data packet captured by Tcpdump from the network generates a unified format record, which contains nine basic characteristics, as shown in Table 1.

Table 1: Basic characteristics table

Feature name	Describe	Data type
Duration	Length of connection	Continuous
Protoles-type	Agreement type, such as tcp, udp, etc.	Discrete type
Servile	Host network services such as http, telnet, etc.	Discrete type
Src-bytes	Data bytes from source host to target host	Continuous
Dst-bytes	Normal or incorrect connection status	Continuous
Flag	Status of error or normal connection	Discrete type
Land	If the connection to the same host or port is 1, the others are 0	Discrete type
Wrong-fragment	Number of faulty sections	Continuous
Urgent	Number of emergency packages	Continuous

3.2 Data preprocessing

Preprocessing is an important data preparation work for the data mining module before mining. According to the requirements of the system algorithm, the data are processed in advance to make the data more in line with the data required by the system mining. Data preprocessing must complete the following tasks:

(1) *Data extraction:* It refers to the integration of operational databases oriented to transaction processing to databases oriented to data mining.

(2) *Data cleaning:* Including some inconsistent data, empty data, data cleaning is to deal with these data, including inconsistent data transformation, vacant data supplement, and so on.

In this work, the two algorithms are used, one is the FP-growth algorithm depicted in Figure 2 and the Apriori algorithm whose flowchart is shown in Figure 3.

The FP-growth algorithm undergoes a process of preparing the clusters and the number of k-clusters are computed followed by the centroid computation. The grouping of different data groups is done depending upon the grouping criteria and the model generation is done using the association rule. Using those association rules from the database, the intrusion scenario is detected.

For the Apriori algorithm, the supports of each item are calculated, and depending upon the support condition, the items are removed or inserted into the frequent item set. The confidence for every non-empty

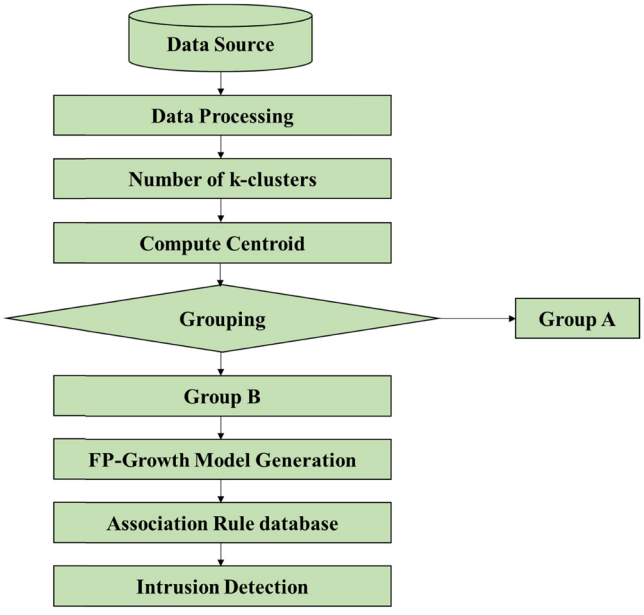

Figure 2: Flowchart of FP-growth algorithm.

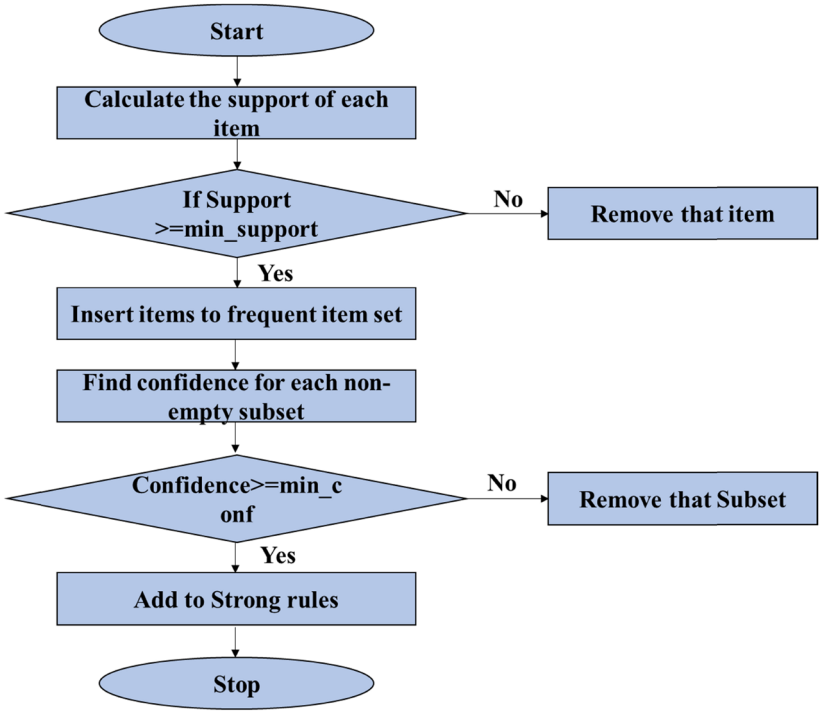

Figure 3: Flowchart of Apriori algorithm.

subset is computed and the selection of the subset is dependent upon the confidence condition. The intrusion detection is accomplished in this algorithm on the basis of the addition of strong rules.

In this experiment, the two algorithms depicted in Figures 2 and 3 based on association rules and clustering analysis are used for experimentation with the subset data kddcup.data_10. percent of KDDcup99. The data are divided into training data and detection data. The training data are used to train the data and build a detection model, and the detection data are used to detect the data and analyze the results of the detection performance [32,33]. When constructing the detection model, it is necessary to filter the abnormal data in the training data, filter out most of the intrusion data and leave only a little intrusion data, and extract the normal data to reach 95%. As long as the detection data are directly extracted as it is, 2,000 network data set records are extracted. The value of the holding degree is set from 0.35 to 8%. The data of 20 mining experiments based on the FP-growth algorithm and the Apriori algorithm are shown in Table 2.

In order to more intuitively reflect the mining effect of the experiment, the conversion is shown in Figure 4.

From the above experiment, it can be found that when the value of the minimum support is smaller, the mining efficiency of the FP-growth algorithm is higher than that of Apriori, and when the value of the minimum support is changed, the mining time is more stable [34]. This is because the FP-growth algorithm does not generate candidate item sets during mining, while the Apriori algorithm generates a large number of candidate item sets, which affects the mining efficiency.

Table 2: Average mining time for different minimum support values

Minimum support (%)	0.35	1.75	3.5	5.25	6.3	7	8
Average time of FP-growth algorithm	4.97	4.77	4.67	4.56	4.44	4.25	4.29
Average time of Apriori algorithm	8.42	8	7.8	7.6	7.4	7.2	7

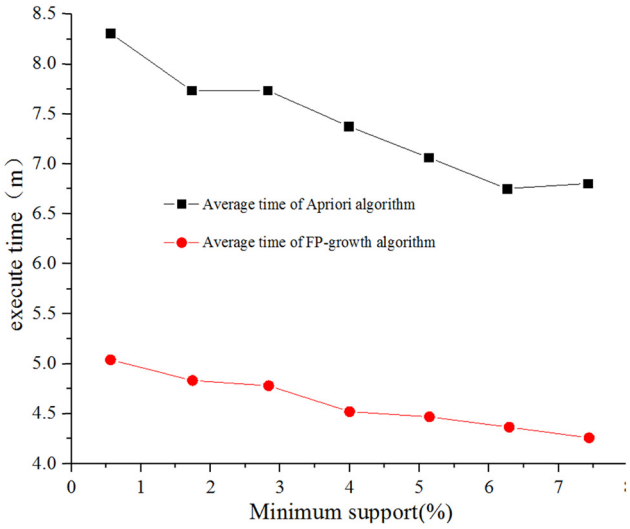

Figure 4: Average mining time of the two algorithms.

The k-means algorithm first needs to input the number of clusters. This number has a greater impact on the results of clustering. Below, we will test the effect of clustering through different numbers of clusters. Select the cluster centers as 10, 12, 16, 18, 20, and 28, respectively, for testing. The training sample data are 8,000 records, and to make the proportion of normal data larger, only 400 intrusion data are left. Reduce 41 attributes to only 13 data field attributes. The two parameters of the improved algorithm are the minimum clustering radius r and exponential factor m. After experiment and observation, the values of the two parameters are $r = 0.3$ and $m = 0.5$, respectively. When $r = 0.3$ and $m = 0.5$, the detection rate and false alarm rate can be selected in a reasonable range. The data of the experimental results are shown in Table 3.

In order to understand the improved performance of the k-means algorithm more intuitively, the experimental results are plotted as shown in Figure 5.

From Figure 5, we can see that the improved k-means algorithm improves the detection rate and reduces the false alarm rate. In short, the improved detection performance of k-means algorithm has been improved.

4 Simulation experimental results and analysis

4.1 Experimental platform

- *Hardware*. The computer is configured with Pentium Core 2.O GHz, DDR memory 2 GB, 7,500-speed notebook hard drive 300 GB.

Table 3: Performance table after algorithm improvement

Number of cluster centers	k-means algorithm		An improved k-means algorithm	
	Detection rate (%)	False alarm rate (%)	Detection rate (%)	False alarm rate (%)
8	43.7	1.3	53.6	1.1
10	52.5	2.2	62.2	2.0
12	68.3	3.7	76.4	2.1
14	77.2	6.5	83.1	3.9
16	82.3	8.6	87.6	5.7

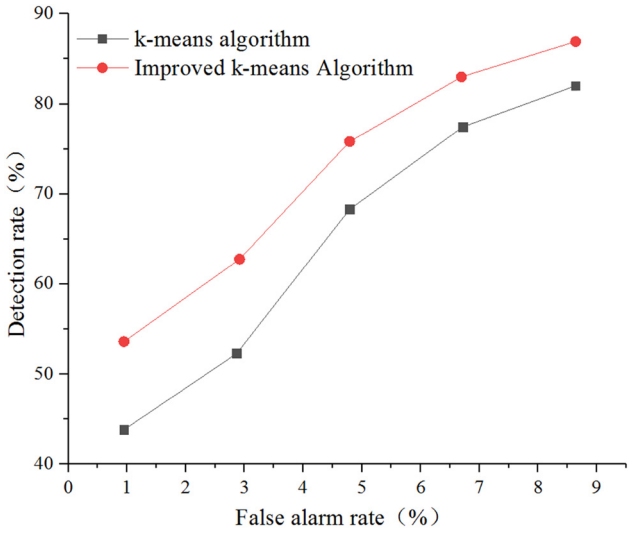

Figure 5: False alarm rate of the two algorithms.

- *Software*. The operating system is Windows XP, the database is SQL 2008, the development tool is C++, the development environment is VC++ 6.0, and the data mining tool is libpcap. V0.8.3.
- *Intrusion detection tools*. The original IDS (Snort v 2. 4 + cluster analysis plug-in + pre-detection plug-in + feature extraction module).
- *Simulate network attack software*. Using IDS Informer v4.0, a simulation attack software developed by Blade is an effective tool for testing IDS.

4.2 Experimental process

- *Data collection*. Use libpcap tool to collect about 50MB data packet samples from the network, calculate the average value and average absolute error of each continuous attribute value, as shown in Table 4, and save it to the Nample.txt file.

Collect the transmitted data packets within 8 h on the network through snort and save the data as normal.log files, which are used to establish the normal behavior pattern of the network training data. IDS Informer is used to send simulated attacks and save the attack data to the attack.log file. There are about 1,000 attack packets in the attack.log file, which are used to test pre-detection and evaluate the effect of the system [35]. In order to reduce unnecessary abusive detection work, the system's pre-detection module will

Table 4: Nample.txt data table

Property name	Average value	Mean absolute error
ip_llen	30560.3	18537.5
ip – tt1	78.03	26.632
tcp-win	0.0023	29431.22
ip_options – len	8.08	0.00268
tcp_options – len	532.8	7.753
dsize	32776.26	576.6
udp – len		13768.6

discard packets that are considered normal. Therefore, the false detection rate can be obtained by calculating the number of packets discarded when detecting the attack.log file.

4.2.1 Experiment of selecting the parameters of the k-means algorithm

Both the cluster analysis module plug-in and the pre-detection module plug-in apply an improved k-means algorithm. Different settings of the clustering radius r exponential factor m will directly affect the effect of the cluster analysis and the judgment result of the pre-detection module.

(1) The influence of clustering radius r (set exponential factor $m = 2$)

When clustering network data packets, the similarity between the network data packet and the center of all network behavior patterns is calculated, and the behavior pattern class to which the data packet belongs is divided. If there is the similarity and clustering radius of a certain network data packet when r is the same, the data packet is classified into the behavior pattern class; otherwise, that is to say, the similarity is greater than all the cluster radius r, indicating that this type of network data packet cannot be classified into any cluster. It is necessary to establish a new cluster for this data packet, and the cluster analysis module performs experiments on the data in the normal.log file. The results are shown in Table 5.

As can be seen from the above table: The clustering radius r will affect the effect of clustering. As the clustering radius r becomes smaller, the network behavior pattern classes obtained increase, and vice versa, the network behavior pattern classes obtained decrease.

Similarly, the clustering radius r also has a certain impact on the false detection rate of the pre-detection module. With the increase of the clustering radius r, the greater the probability of mistaking the attack data as normal data, the higher the false detection rate; otherwise, the lower the false detection rate. Experiment with the data in attack.log through the pre-detection module, and get the false detection rate, as shown in Table 6.

It can be seen from the above table that the larger the clustering radius r, the higher the false detection rate of the pre-detection module. The high false detection rate means that the pre-detection module mistakenly treats some attack packets as normal packets, which leads to detection.

(2) The influence of exponential factor (set cluster radius $r = 4$)

Only when the number of data packets in the abnormal network behavior pattern exceeds the exponential factor, can it become the normal network behavior pattern. This shows that when the exponential factor is smaller, there are fewer abnormal behavior patterns in the network, and there are relatively more normal behavior patterns in the network; conversely, the more abnormal behavior

Table 5: Normal data clustering table

Cluster radius r	Network normal behavioral patterns	Network abnormal behavioral patterns
1	0	148
3	3	132
8	9	92
10	12	76
13	6	56
20	2	27

Table 6: False detection rate results

Cluster radius r	1	3	8	10	15	20
Misdetection rate (%)	0	0.106	23.68	60.8	84.36	97.02

Table 7: Cluster analysis data

Exponential factor *m*	Network normal behavioral patterns	Network abnormal behavioral patterns
1	58	50
2	64	44
4	75	36
6	90	11
8	107	7

patterns in the network, the fewer normal behavior patterns. The cluster analysis module performs experiments on the data in the normal log, and the results are shown in Table 7.

As can be seen from the above table, when the amount of data packets in the abnormal network behavior pattern class reaches the exponential factor, the system will put this type in the normal network behavior pattern library and no longer consider it as an abnormal network behavior pattern. It can be seen that the smaller the exponential factor, the higher the false detection rate.

4.3 Result analysis

Based on the above experiments, it can be seen that the clustering radius and exponential factor will affect the conclusion of the pre-detection module and the effect of clustering. In practical applications, the clustering radius and exponential factor are appropriately adjusted according to the needs to make the detection result better. Usually, the exponential factor is set to a large point, which can avoid taking part of the intrusion operation as normal behavior, and the clustering radius should be adjusted according to the specific situation.

(1) Experiment with unknown intrusion detection capabilities

The feature extraction module is used to generate and extract association rules. From the recorded logs, the improved Apriori algorithm is used to mine frequent item sets, generate association rules and convert them into intrusion detection rules conforming to the system association rule grammar, and add them to the rule base, so that the detection module has the ability to detect new intrusions.

(2) Experiment to verify the detection efficiency of the new system

The new system adds cluster analysis module plug-in, pre-detection module plug-in, and feature extraction module plug-in based on the original system. The improved algorithm is applied in cluster analysis and feature extraction, which improves the detection efficiency of the system. However, since the operation of the added plug-in itself also consumes time, the detection efficiency of the detection system has a certain impact.

The following experiments verify the influence of the parameters of the k-means algorithm on the detection efficiency of the new system.

(1) The influence of clustering radius r (set exponential factor $m = 8$)

Based on the normal network behavior patterns that have been established by the clustering analysis module, we conducted an experiment on the detection efficiency of the attack.log file, and the results are shown in Table 8.

It can be concluded from the above table that the detection time decreases with the increase of the cluster radius r, that is, inversely proportional; that is to say, with the increase of the cluster radius, the detection time becomes shorter. This is because as the cluster radius becomes smaller, the number of clusters generated will increase, and the cluster analysis module will take more time.

(2) The influence of exponential factor (set cluster radius $r = 4$)

Table 8: Detection efficiency of attack.log file

Cluster radius *r*	1	3	8	10	15	20
Detection time *T* (s)	34.5	25.2	16.3	10.4	7.5	6.3

Table 9: Detection efficiency of normal files

Exponential factor *m*	1	2	4	6	8
Interval rate (s)	6.23	7.6	9.06	12.67	14.34

The normal network behavior pattern class is established in the module, and the attack.log file is used to test the efficiency of the experiment. The results are shown in Table 9.

It can be seen from the above table that the time required for system detection is directly proportional to the exponential factor, indicating that the system detection time increases as the exponential factor increases. Because the larger the exponential factor, the time required for the system to establish normal network behavior patterns will also increase, which affects the detection time of the system.

Using KDDCUP99, 10% data set includes a large number of different types of attack data. There are more DOS attack data and 537 data are randomly selected; PROBE attack data are randomly selected, 3,100 data; U2R attack data are less, a total of 60 attack data, 2,124 pieces of R2L data and 2,300 pieces of new attack data. The IDS using the k-means algorithm and the improved k-means algorithm was detected, and the intrusion attacks were run three times. The three detection results are averaged, and the comparison of detection accuracy results is shown in Table 10.

From Table 10, it is revealed that the improved k-means algorithm proposed in this work yields an increased accuracy for intrusion detection from all types of attack evaluated. 7.10% improvement is seen from the DOS type attack, and similarly, 4.63, 78.04, 20.11, and 90.57% improvement is seen for PRONE, R2L, UZR, and new attack types, respectively.

For the clustering analysis module plug-in, the main influencing parameters are the clustering radius *r* and the exponential factor. The setting of its value will have an impact on the clustering effect, and in the pre-detection module, the main impact is on its detection efficiency. In the feature extraction module, the influencing parameter is the support and confidence of the Apriori algorithm. Experiments show that the new system improves the detection efficiency and reduces the false detection rate. It can detect unknown attack behaviors, which shows that it has the ability to deal with new and unknown attacks and also shows its practical application value.

Improve on the basis of analyzing the commonly used k-means algorithm and Apriori algorithm, and apply them to the new data mining IDS. By implementing the proof algorithm, the detection efficiency of the

Table 10: Detection efficiency of the two algorithms

Type of attack	Algorithm	
	k-means algorithm (detection accuracy %)	Improved k-means algorithm (detection accuracy %)
DOS	73.35	78.56
PRONE	72.30	75.65
R2L	35.52	63.24
UZR	60.33	72.46
New attack type	30.12	57.40

system is improved and unknown intrusion attacks can be detected, resulting in reduced false detection rate of the system.

5 Conclusion

k-means algorithm is an initial value-sensitive algorithm in which different k values produce different clustering results. In view of the shortcomings of k-means algorithm, an improved k-means algorithm is proposed in this work. The k value does not need to be determined in advance, and the data set automatically produces an optimal value through clustering. The clustering center does not need to be changed after determination and the whole data set only needs to be scanned once. The efficiency of both the improved k-means algorithm and the clustering effect has improved greatly.

(1) The improved k-means algorithm is verified through experiments. The experimental data show that the improved k-means algorithm has detection efficiency and detection ability. Compared with the previous algorithm, the algorithm improves the detection efficiency and detection ability of the system. A maximum improvement in the detection accuracy of 90.57% can be seen for new attack type intrusion using the proposed algorithm.
(2) Apply the improved algorithm to the IDS and design an IDS based on data mining, which mainly includes pre-detection modules, feature extraction modules, etc. and conducts intrusion detection experiments on the new IDS. The experimental results show that the proposed system improves the detection efficiency and reduces the false detection rate.

The future work in this field will address the shortcomings in the research of IDS based on data mining technology, such as the high cost of the improved Apriori algorithm in space.

Conflict of interest: Authors state no conflict of interest.

References

[1] Yao H, Wang Q, Wang L, Zhang P, Li M, Liu Y. An intrusion detection framework based on hybrid multi-level data mining. Int J Parallel Program. 2019;47(4):740–58.
[2] Salo F, Injadat MN, Nassif AB, Shami A, Essex A. Data mining techniques in intrusion detection systems: a systematic literature review. IEEE Access. 2018;6(1):56046–58.
[3] Olorunnimbe MK, Viktor HL, Paquet E. Dynamic adaptation of online ensembles for drifting data streams. J Intell Inf Syst. 2018;50(2):291–313.
[4] Rathee G, Sharma A, Kumar R, Iqbal R. A secure communicating things network framework for industrial IoT using blockchain technology. Ad Hoc Netw. 2019;94:101933.
[5] Rathee G, Sharma A, Saini H, Kumar R, Iqbal R. A hybrid framework for multimedia data processing in IoT-healthcare using blockchain technology. Multimed Tools Appl. 2020;79:9711–33.
[6] Anderson JP. Computer security threat monitoring and surveillance. Technical report. Fort Washington: James P. Anderson Company; 1980.
[7] Sharma A, Kumar R. An optimal routing scheme for critical healthcare HTH services – an IOT perspective. 2017 Fourth International Conference on Image Information Processing (ICIIP). IEEE; 2017 Dec. p. 1–5.
[8] Sharma A, Tomar R, Chilamkurti N, Kim BG. Blockchain based smart contracts for Internet of medical things in e-healthcare. Electronics. 2020;9(10):1609.
[9] Wael A, Michal Z, Khalid A, Roobaea R, Mehedi M. Mitigation of distributed denial of service attacks in the cloud. Cybern Inf Technol. 2017;17(14):32–5.
[10] Lappas T, Pelechrinis K. Data mining techniques for (network) intrusion detection systems. Riverside CA, 92521: Department of Computer Science and Engineering UC Riverside; 2007.
[11] Deng L, Li D, Yao X, Cox D, Wang H. Mobile network intrusion detection for iot system based on transfer learning algorithm. Clust Comput. 2019;22(4):9889–904.

[12] Zhang J. Detection of network protection security vulnerability intrusion based on data mining. Int J Netw Secur. 2019;21(6):979–84.
[13] Bagaa M, Taleb T, Bernabe JB, Skarmeta A. A machine learning security framework for iot systems. IEEE Access. 2020;8(99):114066–77.
[14] Ektefa M, Memar S, Sidi F, Affendey LS. Intrusion detection using data mining techniques. 2010 International conference on information retrieval & knowledge management (CAMP). IEEE; 2010 Mar. p. 200–3.
[15] Holden N, Freitas AA. A hybrid PSO/ACO algorithm for discovering classification rules in data mining. J Artif Evol Appl. 2008;2008(316145):1–11.
[16] Ardjani F, Sadouni K, Benyettou M. Optimization of SVM multiclass by particle swarm (PSO-SVM). 2010 2nd International Workshop on Database Technology and Applications. IEEE; 2010 Nov. p. 1–4.
[17] Kalaivani S, Vikram A, Gopinath G. An effective swarm optimization based intrusion detection classifier system for cloud computing. 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). 2019 Mar. p.185–8.
[18] Wael A, Michal Z, Khalid A, Roobaea R, Mehedi M. Economic denial of sustainability attacks mitigation in the cloud. Int J Commun Netw Inf Secur. 2017;9(3):420–31.
[19] Petrussenko D. Incrementally learning rules for anomaly detection. Doctoral dissertation. Florida Institute of Technology. Melbourne, Florida; 2009.
[20] Mahoney MV. A machine learning approach to detecting attacks by identifying anomalies in network traffic. Melbourne, Florida; 2003.
[21] Mahoney MV, Chan PK. PHAD: packet header anomaly detection for identifying hostile network traffic. Melbourne, Florida; 2001.
[22] Xiang G, Min W. Applying Semi-supervised cluster algorithm for anomaly detection. 2010 Third international symposium on information processing. IEEE; 2010 Oct. p. 43–45.
[23] Wang Q, Megalooikonomou V. A clustering algorithm for intrusion detection. Data mining, intrusion detection, information assurance, and data networks security 2005. Vol. 5812. International Society for Optics and Photonics; 2005 Mar. p. 31–38.
[24] Mao CH, Lee HM, Parikh D, Chen T, Huang SY. Semi-supervised co-training and active learning based approach for multi-view intrusion detection. Proceedings of the 2009 ACM symposium on Applied Computing; 2009 Mar. p. 2042–8.
[25] Chiu CY, Lee YJ, Chang CC, Luo WY, Huang HC. Semi-supervised learning for false alarm reduction. Industrial conference on data mining. Berlin, Heidelberg: Springer; 2010 July. p. 595–605.
[26] Li J, Zhang W, Li K. A novel semi-supervised SVM based on tri-training for intrusition detection. JCP. 2010;5(4):638–45.
[27] Bhuyan MH, Bhattacharyya DK, Kalita JK. An effective unsupervised network anomaly detection method. Proceedings of the international conference on advances in computing, communications and informatics; 2012 Aug. p. 533–9.
[28] Lane T. A decision-theoritic, semi-supervised model for intrusion detection. Machine learning and data mining for computer security. London: Springer; 2006. p. 157–77.
[29] Fu Z, Papatriantafilou M. Off the wall: lightweight distributed filtering to mitigate distributed denial of service attacks. 2012 IEEE 31st symposium on reliable distributed systems. IEEE; 2012 Oct. p. 207–12.
[30] Fu Z, Papatriantafilou M, Tsigas P. Club: a cluster based framework for mitigating distributed denial of service attacks. Proceedings of the 2011 ACM symposium on applied computing; 2011 Mar. p. 520–7.
[31] Feng J, Shi J, Gao L, Huang H. Application of wireless positioning technology in risk management and control of substation operation site. J Phys Conf Ser. 2020;1544(1):012093 (6pp).
[32] Hong H, Tsangaratos P, Ilia I, Liu J, Zhu AX, Chen W. Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of poyang county, China. Sci Total Environ. 2018;625(Jun 1):575–88.
[33] Panigrahi BK, Das S, Nath TK, Senapati MR. An application of data mining techniques for flood forecasting: application in rivers Daya and Bhargavi, India. J Inst Eng. 2018;99(4):331–42.
[34] Mehedi M, Shamim H. Secure data-exchange protocol in a cloud-based collaborative health care environment. Multimed Tools Appl. 2020;77(9):11121–35.
[35] Sadiq AS, Alkazemi B, Mirjalili S, Noraziah A, Khan S, Ali I, et al. An efficient ids using hybrid magnetic swarm optimization in wanets. IEEE Access. 2018;6:29041–53.