Network Security Threat Detection under Big Data by Using Machine Learning

Jinbao He, Jie Yang, Kangjian Ren, Wenjing Zhang, and Guiquan Li
(Corresponding author: Jinbao He)

Qian'an College, North China University of Science and Technology
Tangshan, Hebei 064400, China
(Email: jinbhe@yeah.net)

(Received Sept. 22, 2018; Revised and Accepted Apr. 11, 2019; First Online July 29, 2019)

Abstract

The efficient detection of network threats has significant meaning to network security. In this study, the problem of network security threat detection by using machine learning was researched. First, K-means clustering algorithm was improved by the stochastic gradient descent, and then it was combined with support vector machine (SVM) algorithm to be used in the tests of the algorithm for different types of network threats. Knowledge Discovery in Database (KDD) 99 data set was also used to test the method in this study. The results showed that the test effects of improved clustering algorithm were obviously better than those of traditional clustering algorithm. Comparing with K-means algorithm, and SVM algorithm, the algorithm in this study had higher detection rate, and lower false alarm rate. Its total detection rate reached 87.1%, and false alarm rate was only 3.1%, which proved the reliability of the algorithm in this study and provided some theoretical support for the further application of algorithm in network security field.

Keywords: Big Data; Machine Learning; Network Security; Network Threat; Support Vector Machine Algorithm

1 Introduction

Under the circumstances of big data, networks, with the development of computer technologies, play an increasingly important role in daily life, and bring great convenience to people's work and life. Meanwhile, network security issues are also becoming more prominent [1]. The rapid development of the Internet provide a lot of network information to the attackers [8], and a large number of network threats have seriously affected the development of the network. Thus, the tests for network threats are obtained more and more concerns and researches [11]. Traditional network threat detection technology has relatively worse detectability, higher false alarm rate, and detection efficiency, but machine learning method can greatly improve these problems.

Machine learning method can efficiently recognize the unknown attack, and has applied on network threat tests [9]. Farnaaz *et al.* [4] applied random forest method in network threat tests, and tested it with NSL-Knowledge Discovery in Database (KDD) data set. The results showed that the method had relatively high detection rate. Haddadpajouh *et al.* [6] used recursive neural network model to test the network threats, and trained the model with data set of 281 malware and 270 benign software. The results showed that the the model had great test effects with its highest detection rate of 98.18%. Chitrakar *et al.* [2] proposed an incremental support vector machine (SVM) algorithm based on candidate support vector. It could be found after comparing with other SVM algorithm that this method had better performance in the network threat tests. Hodo *et al.* [7] analyzed the threats of the Internet of things, and proposed a supervised artificial neural network method to test distributed denial of service attacks. It was found through simulation experiment that the method had an accuracy rate of 99.4%.

Machine learning method can make the network threat detection technology develop in the direction of intelligence, and improve the detection efficiency even in the big data environment, which has a significant positive effect on improving network security. In this study, the clustering algorithm and SVM algorithm in the machine learning were combined with improved clustering algorithm and SVM algorithm to propose a new method of network security threat tests, and the method was tested with KDD 99 data set to prove its reliability, which was beneficial to the further development and application on network threat tests.

2 Network Security Threat and Detection Method

2.1 Network Security Threat under Big Data

With the rapid development of network and the increasing popularity rate of the Internet, the emergence all kinds of network threats events improved the attention to the network security issues. Under the circumstances of big data, the category and number of network threats both obviously increase. Network threats not only expose personal privacy information [5], but also steal and destroy the network information of enterprise data and government agency, which will cause serious consequences. The reasons of network threats may be Transmission Control Protocol / Internet Protocol (TCP/IP) has defects [3], security vulnerabilities of computer software, non-comprehensive network management, hacker attacks, etc.

Current network security technologies include firewalls, data encryption, identity authentication, secure routing, etc. [10], but these are passive network security protection methods that are difficult to effectively deal with in the face of active network threat attacks. Thus, it is necessary to timely and effectively detect network threats and determine whether there are network threats by collecting and analyzing information data in the network, thereby realizing dynamic protection of the network and improving the defense capability of the network. In this study, the cluster method was used to analyze the network threats first.

2.2 K-means Clustering Algorithm

K-means clustering algorithm can divide the date set into different categories. It is assumed that the data set is $S = \{x_1, x_2, \cdots, x_n\}$, $n \in N$, and the number of cluster is K. The specific steps of the algorithm are as in Algorithm 1.

Algorithm 1 K-means clustering algorithm

```
Begin
$k$ initial cluster centers $z_1, z_2, \cdots, z_k$ are randomly selected.
According to initial cluster centers, the samples are divided into $K$ categories as $\{c_1, c_2, \cdots, c_k\}$, and points are divided into cluster $c_i$ of the nearest cluster center.
Cluster centers are recalculated.
if clusters converge then
    The classification will end
else
    It will be classified again
end if
Cluster results are output
End
```

K-means algorithm has good performance and high speed in dealing with big data set, but the initial value in K-means cluster algorithm is sensitive and prone to being trapped in local optimum, which should be improved further.

3 Improved K-means Clustering Algorithm

The clustering algorithm is improved by the method of stochastic gradient descent. It is assumed that $A(\theta)$ is a function that needs to be fitted, $B(\theta)$ is loss function,

$$\begin{aligned} A(\theta) &= \sum_{j=0}^{n} \theta_j x_j \\ B(\theta) &= \frac{1}{2m} \sum_{j=0}^{m} (y^i - A_\theta(x^i))^2 \end{aligned}$$

where θ is the value of the iterative solution, j is the number of parameters, and m is the number of training sets.

In order to control the convergence speed in the calculation process, a learning rate ϵ needs to be set. It is assumed that the present sample is x_k, the search direction is d_k, and $f(\epsilon) = A(x_k + \epsilon d_k)$, $\epsilon > 0$ can be obtained. When $\epsilon = 0$, $f(0) = A(x_k)$, which means $\bigtriangledown f(\epsilon) = \bigtriangledown A(x_k + \epsilon d_k)^T d_k$.

During the gradient descent, the minimum value of $f(\epsilon)$ needs to be found, which is $\epsilon = \arg\min_{\epsilon>0} f(\epsilon) = \arg\min_{\epsilon>0} A(x_k + \epsilon d_k)$, where local minimum value needs to be satisfied as $f(0) = \bigtriangledown A(x_k + \epsilon d_k)^T d_k = 0$. When $\epsilon = 0$, $f'(\epsilon) = \bigtriangledown A(x_k)^T d_k$, and the gradient descent direction is negative gradient $d_k = -\bigtriangledown A(x_k)$. At this time, there must be a ϵ which can obtain $f'(\epsilon) > 0$, where ϵ' is the learning rate that is needed.

The improved clustering algorithm flow is shown in Figure 1.

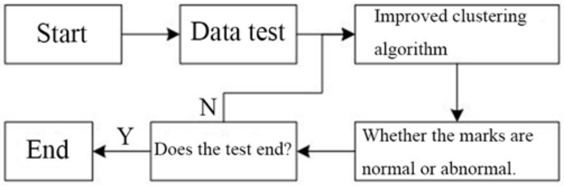
Start
Data test
Improved clustering algorithm
N
End
Y
Does the test end?
Whether the marks are normal or abnormal.

Figure 1: The improved clustering algorithm flow

First, k initial cluster centers are manually determined, then the closest cluster center is found for each sample data x_i and moved toward x_i. In every movement, the learning rate is multiplied constantly till all the samples are divided. Then, the cluster centers are updated, and the process mentioned above will be repeated again till the places of the cluster centers are fixed. The situation whether the clusters are normal or abnormal will be judged, and the test ends.

3.1 Support Vector Machine Algorithm

In order to further improve the detection effect and judge the category of network threats, a new threat detection method based on the combination of the network threat detection method of improved clustering algorithm and SVM is obtained.

The SVM algorithm is a typical dichotomy algorithm [12]. The specific algorithm is as follows:

1) It is assumed that there are samples $\{(x_i, y_i), i = 1, 2, \cdots, l\}$, linear discriminant function is $g(x) = wx + b$, and classification hyperplane is $wx + b = 0$, where w is the direction vector, b is the offset, and $\frac{2}{||w||}$ is class interval. The issue to find optimal separate hyperplane can be represented as $\min \frac{1}{2}||w||^2$, and s.t. $y_i(wx_i + b) - 1 \geq 0$.

2) By Lagrange method, they can be transformed into that when

$$\begin{aligned}\sum_{i=1}^{n} y_i\alpha_i &= 0, \alpha_i \geq 0, i = 1, 2, \cdots, n \\ F(\alpha) &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j (x_i x_j)\end{aligned}$$

is solved, where α is Lagrange multipliers.

3) After solving the samples,

$$f(x) = sgn(wx + b) = sgn\{\sum_{i=1}^{n} \alpha_i y_i (x_i x) + b\}$$

can be obtained.

4) In the current inseparable situation, a slack variable $\xi \geq 0$ is added and a penalty coefficient C is introduced. The objective function is

$$\min \frac{||w||^2}{2} + C\sum_{i=1}^{N} \xi_i y_i[(wx_i) + b] \geq 1 - \xi, \quad i = 1, 2, \cdots, N, \xi \geq 0.$$

5) After solving Lagrange multiplier,

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

can be obtained, where $K(x_i, x_j)$ is kernel function.

6) At last, classification function is

$$f(x) = sgn(\sum_{i=1}^{n} \alpha_i y_i K(x_i x) + b).$$

3.2 Network Threat Detection Method Based on Cluster and SVM

The current common network threats can be divided into four main categories:

Denial of Service Attack (DoS): The use of reasonable service requests to consume network resources, resulting in network overload to stop providing normal services.

Remote to Local (R2L): The remote user uses the vulnerability of the application protocol to send data packets to the target machine and illegally obtain account rights.

User to Root (U2R): Users use system vulnerabilities to enable ordinary accounts to obtain super user privileges.

Probe attack (PROBE): The network is scanned to obtain information such as the IP address.

It can be found that network threat is a multi-classification problem, including four types of threats (DoS, R2L, U2R, Probe) and normal (Normal). These five categories can be split into several dichotomy tasks to train SVM classifier. The multi-classification process is shown in Figure 2.

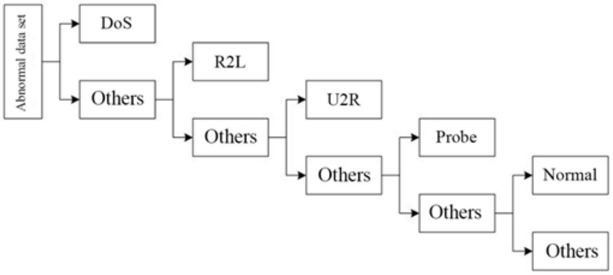

Figure 2: Multi-classification process of network threats

A network threat detection method based on cluster and SVM can be obtained by combining the SVM multi-classification with the network threat detection method based on the improved clustering algorithm in the third section. The specific process is shown in Figure 3.

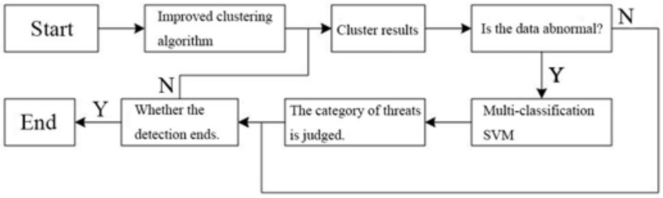
Start
Improved clustering algorithm
Cluster results
Is the data abnormal?
N
Y
Multi-classification SVM
The category of threats is judged.
Whether the detection ends.
N
Y
End

Figure 3: Network threat detection process

Firstly, the improved clustering algorithm is used to divide the data set into different clusters, and the data is judged normally or abnormally. Then the SVM method is used to further classify the abnormal data set to judge the category of network threats. This method can effectively reduce the false alarm rate and improve the detection accuracy of the algorithm.

Table 1: Experimental data

Category	Training set 1	Training set 2	Test set 1	Test set 2
Normal	8800	7500	7000	6800
DoS	570	450	350	400
R2L	200	170	190	140
U2R	150	140	110	120
Probe	170	150	160	120

4 Experimental Analysis

The experimental data came from the KDD 99 data set which consisted of about 5 million training sets and 2 million test sets, including four categories of DoS, R2L, U2R and Probe. 10% of them were selected and divided into four groups. The first and second groups were used as training sets, and the third and fourth groups were used as test sets. The specific information is as shown in Table 1.

Firstly, the performance of the improved K-means clustering algorithm was analyzed, and the traditional clustering algorithm and the improved clustering algorithm were respectively used to detect the network threats. The results of the two methods for the data set are shown in Table 2.

From Table 2 it could be found that the detection effects of the improved clustering were obviously better that those of traditional algorithm. First, from the perspective of the number of clusters, the detection rate and false alarm rate of both algorithms increased as the number of clusters increased. This might be because when the number of clusters was small, normal data was rarely divided into abnormal data, but it was easy for abnormal data to be divided into normal data, which resulted in low detection rate and false alarm rate of the algorithm. When the number of clusters was large, the abnormal data was well divided, but at the same time, many normal data were divided into abnormal data. Thus, the detection rate and false positive rate both increased. When the number of clusters was 25, the detection rate of the improved cluster algorithm was 77.9%, and the false alarm rate was 0.71%, at this time, the detection rate of traditional cluster algorithm was only 54.7%, and the false alarm rate was 0.98%, which proved the reliability of the improved cluster algorithm.

K-means clustering algorithm, SVM algorithm and clustering and SVM-based algorithms proposed in this study were used for network threat detection. The detection results are shown in Figure 4.

From Figure 4 it could be found that the detection effects of the method in this study was obviously better than those of other two single algorithms. Firstly, from perspective of the detection rate, the rate of clustering algorithm was 76.4%, that of SVM algorithm was 81.6%, and that of the algorithm in the study reached 87.1%; from the perspective of false alarm rate, the rate of clustering algorithm was 7.6.%, that of SVM algorithm was 6.3%, and that of the algorithm in this study was 3.1%, which was significantly lower than the rate of cluster algorithm and SVM algorithm, which proved the reliability of the algorithm. The detection results of different network threats used the algorithm in this study are shown in Table 3.

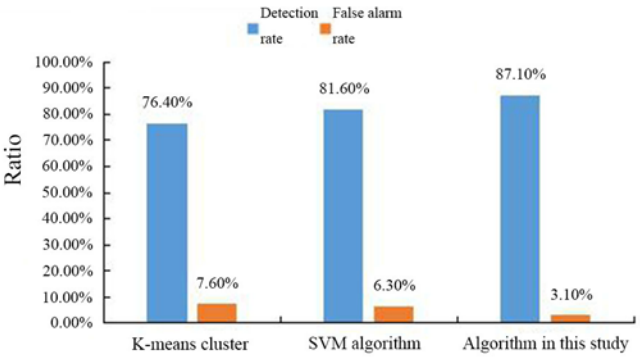

Figure 4: Comparison of different algorithms

From Table 3, it could be found that in the network threat detection, the detection effects of Normal and DoS were better, the detection rates of which were individually 97.6% and 94.5%, while the detection rates of U2R and Probe were lower, and the detection rate of U2R was 78.1%. This might be due to the small amount of data of U2R and Probe, which was not ideal for the classification of these two threats during training, and made the detection effects relatively worse.

5 Discussion

With the development of science and technology and the Internet technology, the categories and quantities of network information data are exploding, and the era of big data starts. Under the circumstances of big data, on the one hand, the network security problem became more and more serious, and the ways of network threats emerge in endlessly, including not only malicious code such as Trojans and worms, but also spyware and advertisement software with unknown content [15]; on the other hand, the network threat detection method can not meet the needs of big data [14], and the detection efficiency is low. Thus, more efficient detection methods are needed. Ma-

Table 2: The test results of the traditional clustering and improved clustering

	Detection rate		**False alarm rate**	
The number of clusters	traditional clustering	improved clustering	traditional clustering	improved clustering
5	18.3%	24.6%	0.41%	0.33%
10	34.7%	37.8%	0.47%	0.41%
15	42.8%	53.6%	0.51%	0.49%
20	51.6%	68.4%	0.72%	0.62%
25	54.7%	77.9%	0.98%	0.71%
30	59.4%	81.4%	1.21%	0.87%

Table 3: Detection results of the algorithm in this study

	Normal	DoS	R2L	U2R	Probe	Total
Detection rate	97.6%	94.5%	86.8%	78.1%	78.5%	87.1%
False alarm rate	2.1%	2.2%	4.3%	2.6%	4.3%	3.1%

chine learning is one of the key technologies of big data processing. It can make the machine learn the law of data through mathematical modeling, and then apply it on similar data. It has a good performance in the processing of massive data, including deep learning [16], decision tree, support vector machine, naive Bayes, clustering algorithm, etc., which have been applied in the field of network security threat detection [13].

In this study, the application of K-means algorithm in network threat detection was researched. In order to improve the shortcomings of clustering algorithm, it was improved by the method of stochastic gradient descent, and then combined with SVM algorithm to obtain a new network threat detection algorithm. According to the experimental results, it could be found that the algorithm designed in this study had a good performance in network threat detection. Firstly, from the perspective of the comparison between traditional clustering algorithm and improved clustering algorithm, the improved clustering algorithm had higher detection rate and lower false alarm rate than those of traditional algorithm, indicating that the improvement of clustering algorithm was effective. Then, in the comparison of K-means clustering algorithm, SVM algorithm and the algorithm in this study in Figure 4, it could be found that compared with the two previous algorithms, the detection effect of the algorithm was significantly higher, the detection rate reached 87.1%, and the false alarm rate was only 3.1%. Finally, from the perspective of the detection effects of different categories of network threats, the algorithm had better detection effects on Normal and DoS, and the detection effects on U2R and Probe was relatively poor, which might be caused by the number of samples.

In order to ensure the network security, only static protection technology cannot provide real-time response to network threats. Dynamic protection technologies are needed to respond proactively to threats inside and outside the network. Network threat detection technology can achieve this. For network threat detection technology, the promotion of popularity and development of the Internet needs higher degrees of intelligence, better detection effects, more secure network, and higher security of the network. It could be found that the network threat detection method based on machine learning algorithm in this study had good reliability and feasibility and was worthy of widespread promotion.

6 Conclusion

In this study, machine learning method was used to propose a network threat detection method based on the combination of improved clustering algorithm and SVM algorithm. From the experimental results, it could be found that the method in this study had 87.1% detection rate and 3.1% false alarm rate, and the detection effects were obviously better than those of single K-means clustering algorithm and SVM algorithm. The detection rate of DoS reached 94.5%, indicating the reliability of the method in this study and providing some theoretical basis for the further application on machine learning algorithm in network threat detection.

References

[1] K. H. Chang, "Security threat assessment of an internet security system using attack tree and vague sets," *The Scientific World Journal*, vol. 2014, pp. 1-9, 2014.

[2] R. Chitrakar, C. Huang, "Selection of candidate support vectors in incremental SVM for network intrusion detection," *Telecom Power Technology*, vol. 45, no. 3, pp. 231-241, 2014.

[3] W. Ding, Z. Yan, R. H. Deng, "A survey on future internet security architectures," *IEEE Access*, vol. 4, pp. 4374-4393, 2016.

[4] N. Farnaaz, M. A. Jabbar, "Random forest modeling for network intrusion detection system," *Procedia Computer Science*, vol. 89, pp. 213-217, 2016.

[5] J. M. Fossaceca, T. A. Mazzuchi, S. Sarkani, "MARK-ELM: Application of a novel multiple kernel learning framework for improving the robustness of network intrusion detection," *Expert Systems with Applications*, vol. 42, no. 8, pp. 4062-4080, 2015.

[6] H. Haddadpajouh, A. Dehghantanha, R. Khayami, K. K. R. Choo, "A deep recurrent neural network based approach for internet of things malware threat hunting," *Future Generation Computer Systems*, vol. 85, pp. 88-96, 2018.

[7] E. Hodo, X. Bellekens, A. Hamilton, *et al.*, "Threat analysis of IoT networks using artificial neural network intrusion detection system," in *IEEE International Symposium on Networks, Computers and Communications (ISNCC'16)*, 2016.

[8] E. D. la Hoz, E. De L. Hoz, A. Ortiz, J. Ortega, B. Prieto, "PCA filtering and probabilistic SOM for network intrusion detection," *Neurocomputing*, vol. 164, pp. 71-81, 2015.

[9] A. J. Malik, W. Shahzad, F. A. Khan, "Hybrid binary PSO and random forests algorithm for network intrusion detection," *Security and Communication Networks*, vol. 8, no. 16, pp. 2646-2660, 2012.

[10] R. Molva, "Internet security architecture," *Computer Networks*, vol. 31, no. 8, pp. 787-804, 2015.

[11] S. Rastegari, P. Hingston, C. P. Lam, "Evolving statistical rulesets for network intrusion detection," *Applied Soft Computing*, vol. 33, pp. 348-359, 2015.

[12] P. Rebentrost, M. Mohseni, S. Lloyd, "Quantum support vector machine for big data classification," *Physical Review Letters*, vol. 113, no. 13, pp. 130503, 2014.

[13] N. Sultana, N. Chilamkurti, W. Peng, R. Alhadad, "Survey on SDN based network intrusion detection system using machine learning approaches," *Peer-to-Peer Networking and Applications*, vol. 12, no. 2, pp. 493-501, 2019.

[14] W. Wang, Y. He, J. Liu, *et al.*, "Constructing important features from massive network traffic for lightweight intrusion detection," *IET Information Security*, vol. 9, no. 6, pp. 374-379, 2015.

[15] L. Zhang, G. B. White, "An approach to detect executable content for anomaly based network intrusion detection," in *IEEE International Parallel and Distributed Processing Symposium*, 2007.

[16] Q. Zhang, L. T. Yang, Z. Chen, *et al.* , "A survey on deep learning for big data," *Information Fusion*, vol. 42, pp. 146-157, 2018.

Biography

Jinbao He, born in November 1987, graduated from North China University of Science and Technology. His research direction is software engineering and integrated technology. He is from Tangshan, Hebei. He has received the master's degree and works as a lecturer in Qian'an College of North China University of Science and Technology. He has published more than 10 papers, including one EI index searching paper, and participated in the editing of a textbook.

Jie Yang, born in June 17, 1985, is a lecturer in Qian'an College of North China University of Science and Technology. He has received the master's degree. He is interested in computer network technology and big data analysis.

Kangjian Ren, born in 1990, has received the master's degree. He is working in Qian'an College of North China University of Science and Technology since June 2014. His research direction is data mining and application.

Wenjing Zhang, born in 1986, has received the master's degree. She is working in Qian'an College of North China University of Science and Technology since August 2012. Her research direction is mechanical engineering.

Guiquan Li graduated from software engineering major in Beijing University of Technology. He participated in the editing of textbook Computer Network Foundation and Fundamentals of Multimedia Technology and has published multiple papers about computer network application.