ICIC Express Letters
Volume **10**, Number **8**, August **2016**

ICIC International ©2016 ISSN 1881-803X
pp. 1957–1964

RESEARCH ON ADAPTIVE DETECTION NETWORK ATTACKS BASED ON THE IMPROVED DATA MINING ALGORITHM IN CLOUD COMPUTING

Guangjun Song[1,2], Bo Li[3] and Xuyang Lu[3]

[1]School of Mathematics, Physics and Information Science
[2]Key Laboratory of Oceanographic Big Data Mining and Application of Zhejiang Province
Zhejiang Ocean University
No. 1, Haida South Road, Lincheng Changzhi Island, Zhoushan 316022, P. R. China
song_gj@126.com

[3]College of Computer and Control Engineering
Qiqihar University
No. 42, Wenhua Street, Qiqihar 161006, P. R. China

Received December 2015; accepted March 2016

Abstract. *Network attacks are one of the biggest threats in cloud security. According to the characteristics of network attacks in cloud computing, an improved k-means algorithm to optimize the clustering number k by function is proposed. The improved k-means algorithm is applied to the intrusion detection model. In the data mining detection module, the improved clustering analysis algorithm is used to adapt and detect those network attacks of different types, which has effectively solved the problems existing in intrusion detection of cloud computing, such as large amount of alarms and high rate of false alarm. Experiment results show that the improved clustering analysis algorithm applied to network attack detection in cloud computing can detect the network attacks efficiently.*

Keywords: Detection technology, Cloud computing, Clustering, Network attacks, Intrusion detection

1. Introduction.

The network attacks have the characteristics of easy launch as well as difficult defense and trace in cloud computing, so it is a hot topic in today's network security field. Denning first proposed IDS (Intrusion Detection System) model [1]. At present, intrusion detection technology has also successively obtained the research and development in cloud computing. Many of the artificial intelligence methods, such as bayesian classifier, hidden markov model, genetic algorithms and artificial neural network, have been applied to a variety of models in IDS [2-6], making IDS protection against network attacks raise a big step in cloud computing. These intrusion detection models based on artificial intelligence are put forward, and their purpose is not only to detect the known attacks but also to find the unknown attacks in network. However, when facing with massive intrusion detection data, these studies have not involved IDS adaptive analysis performance. There is dynamic complexity in cloud computing, the traditional means of intrusion detection cannot meet the demand for cloud environment detection any longer, and as to detection capability or response speed, there are many restrictions. Many traditional IDSs often only apply to small data processing. With the calculation amount increasing, the calculation speed is slowed down greatly, which cannot even be run in could computing.

Data mining has many advantages, such as revealing the rule, extracting the rule, clustering analysis as a data mining algorithm with scalability, dealing with data according to the different types of attributes, and clustering under the constraint condition, which is more suitable for the pattern classification of the large scale data flow. K-means clustering algorithms have fast convergence rate, low complexity and application to large-scale data analysis, so they are used in intrusion detection in cloud environments in a variety of ways [7-11]. However, k-means algorithm is sensitive to the selected initial cluster center, and different initial values often lead to different clustering results; so there is not good stability, and it is easy to fall into local optimum. In order to solve these problems, many improvement programs have been proposed, and applied to intrusion detection models in cloud computing [12-16]. As these improved k-means algorithms have defects of complex parameters, clustering number inaccuracy and so on, the decrease of the quality of the clustering is caused, and some intrusion behaviors are missing in the intrusion detection. For the above problems, an improved k-means algorithm to optimize the clustering numbers k by function is proposed in the paper, and thus the accuracy of the clustering analysis results is greatly improved, and then combined with the characteristics of cloud computing. An intrusion detection model based on the improved k-means algorithm in cloud computing is designed. This model has the advantage of traditional intrusion detection system, and apriori algorithm[17] and the improved k-means algorithm is applied to this model. And the data mining detection module which is more suitable for adaptive detecting network attacks in cloud computing, is also added to this model. Simulation results show that the model has more superior detection performance.

The paper is organized as follows. The improved k-means algorithm and the process of intrusion detection model in cloud computing are discussed in Section 2. The results of simulation and analysis are presented in Section 3. In Section 4, conclusions are given, and further study in the future is pointed out.

2. **Improved Data Mining Detection Model.**

2.1. **Apriori algorithm.**

Association rule mining is one of the technologies most widely used in data mining. Association rule analysis is the method which finds all the support degrees and credibility degrees that exceed the prescribed threshold. Its process is mainly divided into two steps: first, identifying all the support degrees no less than user-specified minimum support degrees threshold itemsets, namely, frequent itemsets; then from the frequent itemsets construct credibility degrees no less than user-specified minimum trusted threshold rules. The classic algorithm to find frequent itemsets is apriori algorithm proposed by Agrawal and Srikant in literature [17].

2.2. **Improved k-means algorithm.**

K-means algorithm is one of the most typical clustering algorithms in data mining, and it can be applied to classify the massive data more quickly and efficiently.

K-means algorithm design processing: first of all, the user determines the exact number of k, and randomly selects k objects, also called the samples; each object is selected as a seed, and it represents a class, also known as the mean or center of the cluster; each object, according to its distance from the center of each cluster, is assigned to the nearest cluster. Then the average value of each cluster is calculated to form a new cluster center, and the process is repeated until the following Formula (1) is convergent.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{1}$$

E is the sum of mean square of all objects in the database, p is the point of space, namely data object, and m_i is the average data of the cluster C_i. According to this criterion, the k clustering itself is as compact as possible and the between clustering is separated as far as possible.

In the traditional k-means algorithm, the number of clustering k should be given by the users in advance. However, in practice, the value of k is difficult to be determined. This is the biggest shortcoming of the algorithm, and also affects the application of the algorithm to some extent. Whether the selection of the number of clusters is reasonable or not is directly decisive to the effect of clustering. In order to solve the above problems effectively, in recent years, many researchers have done a lot of researches in this field. In this paper, a new method is proposed to solve these problems by using the optimization method based on the distance value function, the corresponding mathematical model is established, and the optimization algorithm of k value with the minimum distance criterion is realized.

Set $X = \{x_1, x_2, \ldots, x_n\}$ as a data set containing n objects, and then the cluster distance L_{out} is the sum of the distances from all the cluster center to all the objects center, and the definition of distance L_{out} is as follows:

$$L_{out} = \sum_{i=1}^{k} |m_i - m| \tag{2}$$

In Equation (2), m_i is the center of the cluster C_i, and m is the center of all the objects, which is the mean value of all the objects, and k is the number of clustering. The intra-cluster distance L_i is the sum of the distances from all the objects in C_i to its cluster center, and the definition of distance L_i is as Formula (3), where p_i is any sample in the cluster C_i.

$$L_i = \sum_{p_i \in C_i} |p_i - m_i| \tag{3}$$

Distance value function $D(k)$ is a test function for the optimal number of clustering, and it is an evaluation function about the number of clustering. $D(k)$ is as follows:

$$D(k) = \left| \frac{\sum\limits_{i=1}^{k} \frac{n_i}{n} L_i}{L_{out}} - 1 \right| \tag{4}$$

In Equation (4), the meaning of L_i, L_{out} and k are the same as Equation (2) and Equation (3), n_i is the number of the objects contained in the C_i, and n is the number of all the objects. Weighted L_i is used to avoid having great impact on $D(k)$, and it also reflects the role of the cluster distance L_{out} to $D(k)$.

In the case of all the values that k may take, the data objects are clustered by the improved k-means algorithm and the values of $D(k)$ are calculated in different k value. When the $D(k)$ value reaches the minimum, the clustering results are optimal, and the optimal clustering number is k_{opt}.

To optimize the number of clusters, first of all, the range of the number of k is determined in order to make the number of clusters in reasonable limits. When the minimum value of k is 1, here the distribution of the samples is uniform and has no obvious differences, so it is taken as the minimum value of 2. Therefore, we adopt the experience rules that the value of k is given from ($k = 2, 3, \ldots, \sqrt{n}$) [18], and the optimal value of k is given as follows:

$$\min\{D(k)\}, \quad k = 2, 3, \ldots, \sqrt{n} \tag{5}$$

2.3. **Process and structure of improved detection model in cloud computing.**

(1) Constitution of model structure

The anomaly detection method is used by the network attacking detection model. Imagine network attacks differ from normal behaviors. The normal data pattern is established by analyzing normal data. Therefore, the abnormal data can be judged by whether the two data models make deviation or not. System structure is as follows in Figure 1.

(2) Working process of improved detection model in cloud computing

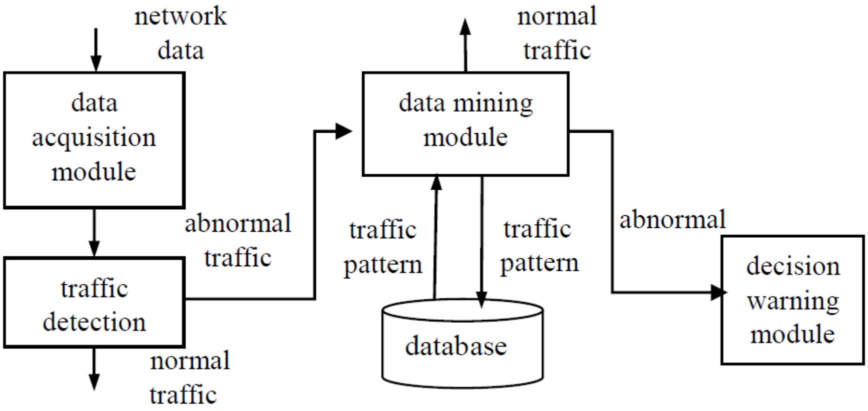

FIGURE 1. The improved model structure of intrusion detection in cloud computing

Use the data acquisition module to collect the packets in the current network in real-time and acquire the current network traffic.

Put current network traffic acquired from the data acquisition module into the traffic detection module and contrast it with the threshold be normal network traffic; if it exceeds the threshold, the network package will be immediately submitted to the data-mining module, and use data mining module to test the submitted abnormal network traffic of packages, through further analysis to testify whether the abnormal network traffic is caused by network attacks. According to the consequence detected by data mining module, if confirmed the current abnormal traffic caused by network attacks, then immediately alarm.

(3) Working process of data-mining module

Use data mining module to detect the abnormal traffic of packets submitted by the traffic detection module. The rules of the abnormal traffic are extracted from the traffic data by apriori algorithm, which can merge network traffic record frequently appearing in time-window into associated record. After the amount of flow records in time-window change into the rules, only several or dozens of records are left in a time-window. The reduction of data in cloud environment can shorten time in data pattern generation and detection, meanwhile increasing the amount of flow data processed in unit time. The rules of the abnormal traffic are judged from whether the traffic patterns are normal or abnormal by the improved k-means algorithm.

After processing in this way, most of the normal data become one or a few large data sets, and the attack data turn into a few small data sets. Attack data differs a good deal with normal data, and it is generally accepted that the clustering with a large amount of data is normal (because in the actual network more than 90% traffic is normal), and the clustering with a small amount of data is abnormal. The normal clustering constitutes the traffic pattern, and these traffic patterns as detection rules are put into the detection module. The number of clustering k in the improved k-means algorithm can be calculated precisely so that the system has better detective accuracy.

Based on the analysis above, the data stream is analyzed by combining apriori algorithm and the improved k-means algorithm. This model not only can detect whether network attacks occur in current network or not, but also can be arranged in other network locations in cloud environment to adapt to monitoring network attacks completely.

3. Simulation Test and Result Analysis.

KDDCUP1999 intrusion detection evaluation datasets established by American MIT Lincoln Laboratory are adopted, which have been widely used in simulation tests of intrusion detection methods [19].

In the model system performance testing, the indicators of DR (Detection Rate), FPR (False Positive Rate) and Accuracy are used to evaluate the performance of test model. Detection rate refers to ratio of the total number of detected attacks and the total number of actual invasion when the system is attacked. False positive rate is the ratio of the total number of misjudged invasion data and the total number of normal data. Accuracy is the ratio of the total number of accurately judged data, and the total number of data in the test sets.

3.1. **The results of clustering algorithm in different threshold.** In the experiments, the test set A, the test set B and the test set C are obtained by three times of random selection from 9000 normal data and 1000 attack data that selected from the *KDDCUP1999* datasets. In order to reach a better detection effect, and to get a higher detection rate and a lower false alarm rate, we choose the model with improved k-means algorithm to experiment the three test sets respectively in different abnormal traffic threshold e.

Table 1. The detection results of the improved model in different threshold e

Threshold	*Test set A*		*Test set B*		*Test set C*	
e(%)	*DR(%)*	*FPR(%)*	*DR(%)*	*FPR(%)*	*DR(%)*	*FPR(%)*
5	68.75	0.23	70.25	0.25	66.95	0.19
10	79.63	0.41	81.76	0.49	77.48	0.38
15	87.56	0.76	89.34	0.81	85.53	0.62
20	94.74	1.89	95.63	2.18	94.67	1.69
25	95.92	3.41	96.82	3.73	95.23	2.97

As Table 1 shows, with the increase of threshold, the detection rate increases gradually and the false alarm rate also increases, and the overall system performance also changes. When the threshold is 20%, the average detection rate of the system is 95.01%, the average false alarm rate is 1.92%, and the overall system performance is relatively optimal at this moment.

3.2. **The improved model test results and analysis.** Respectively, the models that use traditional k-means and the improved k-means algorithm are tested with the detection rate and the false positive rate, and the threshold is taken for 20%. The test results are as shown in Table 2.

Table 2. The comparison results of the models

Test set	*k-means*		*Improved k-means*	
	DR(%)	*FPR(%)*	*DR(%)*	*FPR(%)*
Test 1	91.32	3.89	94.26	1.98
Test 2	91.65	4.36	95.93	1.61
Test 3	90.93	5.63	95.37	2.19

As can be seen from Table 2, under the same conditions, the detection rate of the improved k-means algorithm model is significantly higher than that of k-means model and the false positive rate is lower.

So 20% is selected as threshold to experiment the three test sets many times, the average of the test results is taken as the detection rate and the false alarm rate of the system, and the detection results of the model that adopted the traditional k-means algorithm and the improved k-means algorithm are contrasted respectively. The comparison results are shown in Figure 2 and Figure 3. And Figure 4 shows the comparing results of the accuracy.

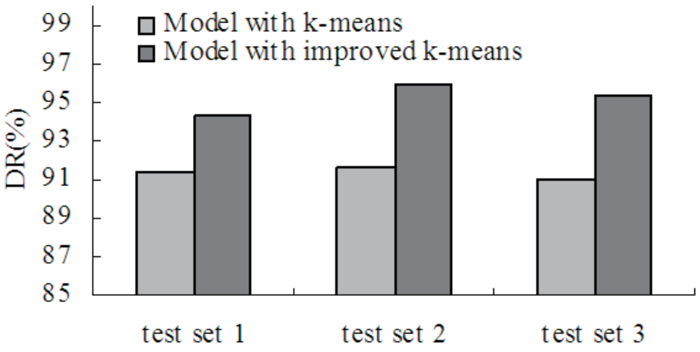

Figure 2. Comparison results of using DR

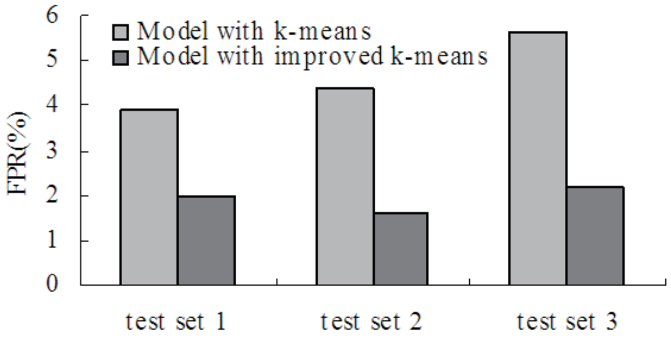

Figure 3. Comparison results of using FPR

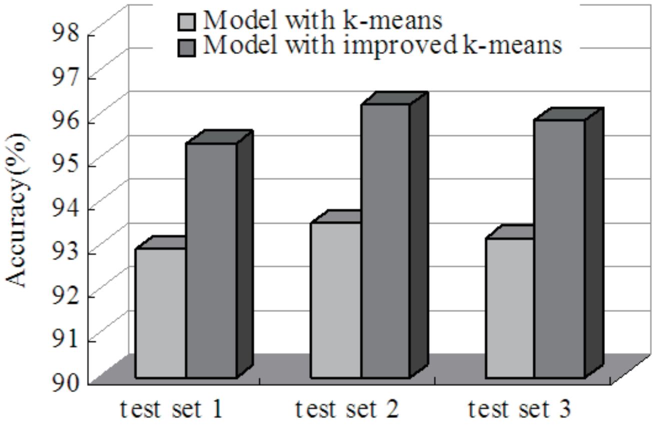

Figure 4. Comparison results of using accuracy

As can be seen from Figure 4, under the same conditions, the accuracy of the improved data mining algorithm model is significantly higher than that of traditional data mining model.

4. **Conclusions.** Based on the characteristics of virtualization and distribution in cloud computing, the paper puts forward an intrusion detection technology that uses the improved k-means algorithm in cloud computing. A network attacks intrusion detection

model based on the improved data mining algorithm is designed. Through the discussion above, conclusions and future work are summarized as follows. First, an improved k-means algorithm to optimize the clustering numbers k by function is proposed. Second, an intrusion detection model suitable for cloud computing is designed, the improved k-means algorithm is applied to the intrusion detection model, and the detection rate of the model system against network attacks is improved while the false alarm rate of the system is reduced. Third, the model combines apriori algorithm with the improved k-means algorithm, which enhances the detection rate and accuracy for the network attacks. The experiments prove that in this paper, the design of intrusion detection model in the cloud computing can effectively defend the network attacks. However, the practicability and real-time performance of the model system are not ideal, and the detection performance also needs to be perfected, which is also our future research direction.

Acknowledgment. This work is supported by the National Natural Science Foundation of Zhejiang Province, China, No. LY16F020014, and the Scientific Research Foundation of Zhejiang Ocean University, China, 2014, No. 21065013814.

REFERENCES

[1] D. Denning, An intrusion detection model, *IEEE Trans. Software Engineering*, vol.13, no.2, pp.222-232, 1987.

[2] C. N. Modi, D. R. Patel et al., Bayesian classifier and Snort based network intrusion detection system in cloud computing, *Proc. of the 3rd IEEE International Conference on Computing Communication and Networking Technologies*, pp.1-7, 2012.

[3] P. Kumar, V. Sehgal, K. Shah et al., A novel approach for security in cloud computing using hidden Markov model and clustering, *Proc. of the 1st World Congress on Information and Communication Technologies*, pp.810-815, 2011.

[4] A. Kannan, G. Q. Maguire Jr, A. Sharma et al., Genetic algorithm based feature selection algorithm for effective intrusion detection in cloud networks, *Proc. of the 12th IEEE International Conference on Data Mining Workshops*, pp.416-423, 2012.

[5] S. T. F. Al-Janabi and H. A. Saeed, A neural network based anomaly intrusion detection system, *Proc. of the 1st IEEE International Conference on Developments in E-systems Engineering*, pp.221-226, 2011.

[6] C. Modi, D. Patel, B. Borisanyia et al., A survey on security issues and solutions at different layers of cloud computing, *Journal of Supercomputing*, vol.63, no.2, pp.561-592, 2013.

[7] L. Xiao, Z. Shao and G. Liu, K-means algorithm based on particle swarm optimization algorithm for anomaly intrusion detection, *Proc. of the 6th IEEE World Congress on Intelligent Control and Automation*, vol.2, pp.5854-5858, 2006.

[8] L. Jing, M. K. Ng and J. Z. Huang, An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data, *IEEE Trans. Knowledge and Data Engineering*, vol.19, no.8, pp.1026-1041, 2007.

[9] H. Gao, D. Zhu and X. Wang, A parallel clustering ensemble algorithm for intrusion detection system, *Proc. of the 9th IEEE International Symposium on Distributed Computing and Applications to Business Engineering and Science*, pp.450-453, 2010.

[10] W. Yassin, N. I. Udzir and Z. Muda, A cloud-based intrusion detection service framework, *Proc. of the International Conference on Cyber Security, Cyber Warfare and Digital Forensic*, pp.213-218, 2012.

[11] A. M. Riad, I. Elhenawy, A. Hassan et al., Visualize network anomaly detection by using k-means clustering algorithm, *International Journal of Computer Networks & Communications*, vol.5, no.5, pp.195-208, 2013.

[12] L. Tian and W. Jianwen, Research on network intrusion detection system based on improved k-means clustering algorithm, *Proc. of the International Forum on Computer Science-Technology and Applications*, pp.76-79, 2009.

[13] S. K. Sharma, P. Pandey and S. K. Tiwari, An improved network intrusion detection technique based on k-means clustering via naïve bayes classification, *Proc. of IEEE International Conference on Advances in Engineering, Science and Management*, pp.417-422, 2012.

[14] L. Bai, J. Y. Liang, C. Sui et al., Fast global k-means clustering based on local geometrical information, *Information Sciences*, vol.245, no.1, pp.168-180, 2013.

[15] E. P. Nikolova and V. G. Jecheva, An adaptive approach of clustering application in the intrusion detection systems, *Open Journal of Information Security and Applications*, vol.1, no.3, pp.1-10, 2014.

[16] R. Brar and N. Sharma, A novel density based k-means clustering algorithm for intrusion detection, *Journal of Network Communications and Emerging Technologies*, vol.3, no.3, pp.17-22, 2015.

[17] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, *Proc. of the 20th International Conference on Very Large Databases*, pp.487-499, 1994.

[18] R. M. Ramze, B. P. F. Lelieveldt and J. H. C. Reiber, A new cluster validity indexes for the fuzzy c-means, *Pattern Recognition Letters*, vol.19, no.1, pp.237-246, 1998.

[19] *KDD Cup 1999 Data*, http://kdd.ics.uci.edu/databases/kddcup99, 1999.