CHEMICAL ENGINEERING TRANSACTIONS
VOL. 51, 2016
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Tichun Wang, Hongyang Zhang, Lei Tian

ISBN 978-88-95608-43-3; **ISSN** 2283-9216
DOI: 10.3303/CET1651102

Application Research on Data Mining Algorithm in Intrusion Detection System

Weizu Wu*[a], Liqun Liu[a], Bing Xu[b]

[a]College of Information,GuangDong Ocean University, ZhanJiang, GuangDong 524088 China
[b]College of Computer and Electronic Information,Guangdong University of Petrochemical Technology, Moming, GuangDong 525000 China
zjwuweizu@163.com

As is known to all, the Internet is open and shared. These two characteristics make the Internet has a wealth of resources. However, they bring security risks to the Internet. One solution is to establish a security system that is relatively easy to achieve, and to establish the appropriate security assistance system in accordance with a certain security policy. In terms of computer security in the network environment, we need a kind of technology that can detect and report the unauthorized or abnormal phenomena in the system, that is, the intrusion detection technology. Data mining is a data analysis and processing technology which is a kind of widely used. Data mining technology can quickly and effectively analyze the big data, and find out the useful information and knowledge. Clustering analysis is an important tool in data mining, and clustering analysis is used to find the potential relationship between the data attributes. The k-means algorithm is a typical clustering algorithm that has the advantages of fast convergence speed and strong local search ability. But k-means algorithm has some defects, such as the sensitivity to the initial centre, easy to fall into local optimum. In order to improve the detection effect of intrusion detection system, this paper researches on the commonly used clustering and classification techniques in data mining. Then, in order to solve the problem that the detection result is affected by the initial clustering centre and number setting, we propose a k-means clustering algorithm based on genetic algorithm. Genetic algorithm has good global optimization ability, and the improved crossover operator and mutation operator can be used to generate a better chromosome. Finally, through the simulation of KDD CUP99 data set, the feasibility and validity of the proposed approach is verified. The experimental results show that this method improves the accuracy of clustering, speeds up the convergence rate, and enhances the stability of the algorithm.

1. Introduction

With the development of the Internet, it has become an important part of human life. Especially after the realization of globalization and information technology, many enterprises, government agencies and individuals carry out a variety of business and services in the open Internet, such as financial enterprises carry out online banking. Network has brought convenience to people's work and life, thus it has changed the way of people's work and live. This meets the needs of people with fast, efficient and convenient life(Zha and Wang, 2010; Qin and Hu, 2002; Wang and Zhao, 2010; Ji, 2007). As is known to all, the Internet is open and shared. These two characteristics make the Internet has a wealth of resource. However, they bring security risks to the Internet. One solution is to establish a security system that is relatively easy to achieve, and to establish the appropriate security assistance system in accordance with a certain security policy. In terms of computer security in the network environment, we need a kind of technology that can detect and report the unauthorized or abnormal phenomena in the system, that is, the intrusion detection technology. According to the different detection methods, it can be divided into two categories that are misuse detection and anomaly detection.

The representative research methods of anomaly detection are expert system method, statistical analysis method, neural network method and computer immunology. The anomaly detection systems based on statistical method are IDES(Lunt, 1989; Anderson, 1995; Porras and Neumann, 1997). The advantage of

Please cite this article as: Wu W.Z., Liu L.Q., Xu B., 2016, Application research on data mining algorithm in intrusion detection system, Chemical Engineering Transactions, 51, 607-612 DOI:10.3303/CET1651102

statistical analysis is that the user's behaviour patterns can be accurately characterized by carefully selected metrics, so as to find behaviours that violate the security policy. One of the characteristics of the neural network is that it has the ability to learn, so some IDS use it to study the behaviour of users. NNID(Jake et.al, 1998) is a neural network intrusion detection system, which reads the user's command line log, uses neural network to learn the user behaviour, and detects the serious deviation in user behaviour. Forrest et.al (1996) explore the problem of computer security in a new perspective, and they find some similarities between the protection mechanism of the computer system and the biological immune system. The most basic ability of the immune system is to recognize the self or non self, which is similar to the concept of anomaly detection in intrusion detection. The representative research methods of misuse detection are pattern matching, state transition analysis, and feature analysis. Pattern matching system is proposed by Kumarl etal(1994), which is one of the most widely used detection methods and mechanisms in the field of intrusion detection. They compare the collected information with the known intrusion data, so as to discover the behaviour contrary to the security policy. Feature analysis uses an intuitive way to convert the semantic description of the attack to the information that can be found in the audit log (Lin et al., 1998). The state transition analysis is used to represent a series of actions that the intruder performs, and these behaviours and requirements are expressed as a state transition graph (Porras and Kemmerer, 1992; Vigna and Kemmerer, 1998).

Clustering analysis method is one of the core technologies in data mining, and its related algorithm is more and more valued by scholars. A lot of clustering algorithm analysis is emerging. In reference (Hu, 2013), the hierarchical k-means clustering tree is used to automatically select the number of clusters to obtain a better clustering effect. Yu Qianqian, Dai Yueming, and Li Jingjing(2013) propose the method of parallel k-means. This method not only has a good speedup and scalability, but also its convergence and clustering accuracy are improved. In reference (Lin et al., 2014), the improved particle swarm optimization algorithm is used to select the initial cluster centre, and then the k-means algorithm is used to optimize the clustering. Finally, according to the condition of multi class merging, the best clustering results are obtained. In addition, a method based on simulated annealing algorithm is proposed by Liu Hanmei, Zhang Peng(2013). It can effectively prevent the k-means algorithm falling into local optimum.

2. An improved K-means clustering algorithm

2.1 Introduction of K-means clustering algorithm

MacQueen proposes the k-means algorithm which is an effective algorithm, and it is widely used in the field of science and industry. The working mechanism of the k-means algorithm is to divide the *n* sample points into *k* clusters, and the sample points in each cluster have high similarity. The degree of similarity between the sample points of each cluster is relatively low, the similarity calculation is based on the average value of sample points in a cluster. Algorithm specific procedures are as follows:

(1)Select the *k* sample points from the sample dataset *D*, and the *k* values are assigned to the initial clustering centre ($\mu_1^1,\mu_2^1,\ldots,\mu_k^1$).

(2)In the first *j* th iteration, calculate the Euclidean distance $d(t,i)$ between all points $p_t(t=1,2,\ldots,n)$ in the sample and the cluster centre μ_i^j in proper sequence.

$$d(t,i)=\sqrt{(p_t-\mu_i^j)^2} \tag{1}$$

(3) Find out the minimum distance of p_t about μ_i^j, put the p_t into the smallest cluster about the distance of μ_i^j,

(4) Update the clustering centre of each cluster.

$$\mu_i^{j+1}=\frac{1}{n_i}\sum_{t=1}^{n_i}p_{it} \tag{2}$$

(5) Calculate the square error E_i of all the points in the data set *D*, and compare with the previous error E_i.

$$E_i=\sum_{i=1}^{k}\sum_{t=1}^{n_i}\left|p_{it}-\mu_i^{j+1}\right| \tag{3}$$

If $\left|E_{i+1}-E_i\right|<\delta$, then the algorithm ends. Otherwise, turn to the step 2.

2.2 Intrusion detection algorithm based on Genetic Improvement

(1) Selection operator

The selection operator reflects the principle of survival of the fittest. The selection operator is used to select individuals with high fitness, and then their excellent genes are inherited to the next generation. The

commonly used selection operators are the proportional selection method, optimal preservation strategy method, expected value method, sorting and selection method, etc. In this paper, the combination of the optimal preservation strategy method and the proportional selection method are used.

(2) Crossover operator and mutation operator

In the traditional adaptive genetic algorithm, the crossover probability and mutation probability are changed with the change of the fitness value. With evolution, they are adapted to change between 0 and 1. But in the following case, genetic algorithm is easy to fall into local optimum. The probability and mutation probability of maximum fitness value of individual crossover are 0. Alternatively, the crossover probability and mutation probability of those individuals which are similar to the greatest fitness values are close to 0. To prevent these problems, an improved adaptive genetic algorithm is proposed in this paper.

The improved crossover probability and mutation probability formula are as follows:

$$p_c = \begin{cases} p_{c_1} - \dfrac{(p_{c_1} - p_{c_2})(f' - f_{\max})}{f_{\max} - f_{avg}} & f' \geq f_{avg} \\ p_{c_1} & f' < f_{avg} \end{cases} \tag{4}$$

$$p_m = \begin{cases} p_{m_1} - \dfrac{(p_{m_1} - p_{m_2})(f - f_{\max})}{f_{\max} - f_{avg}} & f \geq f_{avg} \\ p_{m_1} & f < f_{avg} \end{cases} \tag{5}$$

Where, f_{avg} is average fitness value of all individuals in the population, f_{max} is the largest individual fitness value in the population, f is the fitness value of the crossover individual with larger fitness value, and f' is the fitness value of variant individual. Then, the value p_{c1} is 0.95, the value of p_{c2} is 0.7, the value of p_{m1} is 0.1, and the value of p_{m1} is 0.01.

Next, we give the operation steps of k-means clustering based on adaptive ant colony improvement.

Step1: Setting parameters.

The number of initial clusters are c, the population size is m, the crossover probability is p_c, and the mutation probability is p_m. The maximum number of iterations is T, and the adaptive parameter is c.

Step2: Randomly generated initial population.

Step3: Calculates the fitness value of each individual in a group.

Step4: Generate a new generation of groups by the operation of selection, crossover, mutation, and k-means.

Step5: Repeat the step 3 and 4 until the maximum number of iterations is reached.

Step6: Calculate the new population fitness value, and use the individual which has the maximum fitness value as the centre, so as to carry out the calculation of k-means clustering.

Step7: Output clustering results.

3. Simulation experiment and result analysis

3.1 The experimental data

In this paper, an improved k-means clustering algorithm is used to monitor the behavior of the network intrusion, and the KDD CUP99 data set is used as the sample data set. Because the CUP KDD sample data set is too large, and the length of this article is limited, we select only 40 records to experiment. The 40 data contains 29 normal records, 2 records are attacked by the Neptune, 7 records are attacked by the Smurf, and 2 records are attacked by the ipsweep. Then, we use genetic algorithm to find the initial center of clustering and the optimal center, as shown in Table 1 and table 2.

3.2 The experiment steps and the result analysis

As we can see from table 2, the 40 records of the 41 attributes have all found their optimal cluster centers. Next, we calculate the final classification based on the k-means algorithm. In order to verify the validity of the method in this paper, we compare with the hierarchical clustering algorithm and k-means clustering algorithm. The final classification results are shown in table 3.

From Figure 1, we can see that compared to the traditional k-means clustering and hierarchical clustering intrusion detection algorithm, the improved k-means clustering algorithm in this paper has a higher classification accuracy. For unknown types of attacks, the improved k-means clustering method based on genetic algorithm has a certain ability of detection. To a certain extent, it overcomes the defect that the rule intrusion detection system can only detect the known attack behaviour. In this paper, the improved k-means clustering algorithm automatically generates the number of clusters, which reduces the dependence of the

algorithm on K value. In addition, the initial cluster centres are obtained through the analysis of the specific data, so that the cluster centres are selected more accurately.

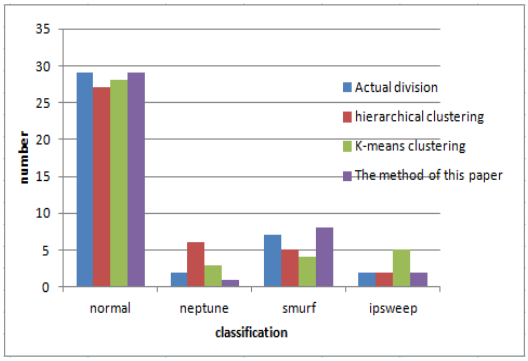

Figure 1. Comparison of classification results of various methods

Table 1: Initial cluster center

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
duration	0	0	0	0
protocol_type	1	0	0	0
service	2	0	0	0
flag	0	0	0	0
src_bytes	1032	199	213	364
dst_bytes	0	45076	4663	11189
land	0	0	0	0
wrong_fragment	0	0	0	0
urgent	0	0	0	0
hot	0	0	0	0
num_failed_logins	0	0	0	0
logged_in	0	1	1	1
num_compromised	0	0	0	0
root_shell	0	0	0	0
su_attempted	0	0	0	0
num_root	0	0	0	0
num_file_creations	0	0	0	0
num_shells	0	0	0	0
num_access_files	0	0	0	0
num_outbound_cmds	0	0	0	0
is_hot_login	0	0	0	0
is_guest_login	0	0	0	0
count	511	1	22	1
srv_count	511	4	24	1
serror_rate	0	0	0	0
srv_serror_rate	0	0	0	0
rerror_rate	0	0	0	0
srv_rerror_rate	0	0	0	0
same_srv_rate	1	1	1	1
diff_srv_rate	0	0	0	0
srv_diff_host_rate	0	0.75	0.12	0
dst_host_count	255	62	255	1
dst_host_srv_count	255	62	255	64
dst_host_same_srv_rate	1	1	1	1
dst_host_diff_srv_rate	0	0	0	0
dst_host_same_src_port_rate	1	0.02	0	1
dst_host_srv_diff_host_rate	0	0	0	0.03
dst_host_serror_rate	0	0	0	0
dst_host_srv_serror_rate	0	0	0	0
dst_host_rerror_rate	0	0	0	0
dst_host_srv_rerror_rate	0	0	0	0

Table 2: Optimal cluster center

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
duration	0	0	0	0
protocol_type	0.310344828	0	0	0
service	0.75862069	0	0	0
flag	0.137931034	0	0	0
src_bytes	293.2413793	199	201.125	296
dst_bytes	362.6896552	45076	3054.875	9892
land	0	0	0	0
wrong_fragment	0	0	0	0
urgent	0	0	0	0
hot	0	0	0	0
num_failed_logins	0	0	0	0
logged_in	0.551724138	1	1	1
num_compromised	0	0	0	0
root_shell	0	0	0	0
su_attempted	0	0	0	0
num_root	0	0	0	0
num_file_creations	0	0	0	0
num_shells	0	0	0	0
num_access_files	0	0	0	0
num_outbound_cmds	0	0	0	0
is_hot_login	0	0	0	0
is_guest_login	0	0	0	0
count	88.5862069	1	18.75	7.5
srv_count	81	4	20.375	7.5
serror_rate	0.137931034	0	0	0
srv_serror_rate	0.137931034	0	0	0
rerror_rate	0	0	0	0
srv_rerror_rate	0	0	0	0
same_srv_rate	0.870344828	1	1	1
diff_srv_rate	0.009310345	0	0	0
srv_diff_host_rate	0.183793103	0.75	0.14625	0
dst_host_count	190.7931034	62	177.125	21
dst_host_srv_count	200.7931034	62	199.875	98
dst_host_same_srv_rate	0.865172414	1	1	1
dst_host_diff_srv_rate	0.009655172	0	0	0
dst_host_same_src_port_rate	0.314137931	0.02	0.01	0.51
dst_host_srv_diff_host_rate	0.073103448	0	0.01	0.035
dst_host_serror_rate	0.137931034	0	0	0
dst_host_srv_serror_rate	0.137931034	0	0	0
dst_host_rerror_rate	0	0	0	0
dst_host_srv_rerror_rate	0	0	0	0

Table 3: Final classification results

	Normal	Neptune	Smurf	Ipsweep
Actual division	29	2	7	2
hierarchical clustering	27	6	5	2
k-means clustering	28	3	4	5
The method of this paper	29	1	8	2

4. Conclusions

In order to improve the detection effect of intrusion detection system, this paper researches on the commonly used clustering and classification techniques in data mining. Then, in order to solve the problem that the detection result is affected by the initial clustering centre and number setting, we propose a k-means clustering algorithm based on genetic algorithm. Genetic algorithm has good global optimization ability, and the improved crossover operator and mutation operator can be used to generate a better chromosome. Finally, through the simulation of KDD CUP99 data set, the feasibility and validity of the proposed approach is verified.

The experimental results show that this method improves the accuracy of clustering, speeds up the convergence rate, and enhances the stability of the algorithm.

References

Anderson D., Lunt T.F., Javitz H., etc, 1995, Detecting unusual program behaviour using the statistical component of the next-generation intrusion detection expert system[C].Technical Report, SRI International, Menlo Park,CA, 5-10.

Forrest S., Hofmeyr S.A., Somayaji A., 1996, A sense of self for unix processes[C].In Proceedings of the 1996 IEEE symposium on security and privacy, 120-128.

Hu W., 2013, Improved hierarchical K mean clustering algorithm [J]. computer engineering and application, 49 (2): 157-159.

Kumar S., Spaffbrd E.H., 1994, A Patten Matching Model for Misuse Intrusion Detection[J]. The COAST Projcet Dep.of Computer Sciences Purdue University.

Lei J., 2007, The research of intrusion detection technology based on Data Mining[J]. Journal of Shanghai Jiao Tong University.

Lin J., Wang X.S., 1998, Abstraction-based misuse detection: High level specification and adaptable strategies[C].In the 1lth Computer Security Foundations Workshop, Rockport, MA, June.

Lin Y.C., Fu Q. et al., 2014, Application of PSO-means clustering algorithm with multi class [J]. computer system, 23 (2): 160-165.

Liu H.M., Zhang P., 2013, Optimization of K-means clustering algorithm based on simulated annealing algorithm [J]. Western China Science and technology, 12 (6): 23-24.

Lunt T.F., 1989, Real-Time Intrusion Detection[C].In compcon'89: IEEE Computer Conference, 2:348-353.

Porras P. and Neumann P.G., 1997, Event monitoring enabling responses to anomalous live disturbances. In the 19th National Information systems Security Conference, Baltimore, MD, 10:353-365.

Porras P., and Kemmerer R., 1992, Penetration state transition analysis-a rule-based intrusion detection approach[C].In 8th Annual Computer Security Conference, 12:220-229.

Qin A.M., Hu C.Z., 2002, Application of data mining technology in network attack detection [J]. computer engineering and application, (11): 177-180.

Ryan J., Lin M.J., Miikkulalnen R., 1998, Intrusion Detection with Neural Networks[M]. Advances in Neural Information Processing Systems 10, Cambridge, MA: MIT Press.

Vigna G. and Kemmerer R., 1998, A Network-based Intrusion Detection Approach[C].In Proceedings of the 14th Anuua Computer Security Application Conference, 10-16.

Wang M.D., Zhao G.H., 2010, High speed intrusion detection algorithm based on network traffic characteristics analysis [J]. computer application research, 27 (9): 3484-3486.

Yu Q.Q., Dai Y.M., Li J.J., 2013, Based on parallel MapReduce ACO-K-means clustering method [J]. Computer engineering and applications, 49 (16) 117-120.

Zha Q.M., Wang R.G., 2010, Intrusion detection method based on quantum genetic clustering [J]. computer application research, 27 (1): 240-243