TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.1, January 2014, pp. 762 ~ 770
DOI: http://dx.doi.org/10.11591/telkomnika.v12i1.3353

Clustering and Hybrid Genetic Algorithm based Intrusion Detection Strategy

Li Liu*[1], Pengyuan Wan[2], Yingmei Wang[3], Songtao Liu[4]
[1]School of Automation & Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2]School of Computer & Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[3]Beijing Graphics Institute, Beijing 100029, China
[4]Beijing AVIC Information Technology Company, Aviation Industry Corporation of China
*Corresponding author email: liuli@ustb.edu.cn

Abstract

Ad hoc networks face serious security threat due to its inherent weaknesses. Intrusion detection is crucial technology in protecting the security of Ad hoc networks. Recently, Intrusion Detection Systems (IDS) face open issues, such as how to make use of intrusion detection technologies to excavate normal/abnormal behaviors from a lot of initialized data and dig out invasion models later for intrusion detection automatically and effectively. In this paper, we propose an enhanced algorithm combined improved clustering algorithm with Hybrid Genetic Algorithm (HGA), called Enhanced Intrusion Detection Algorithm (EIDA) for intrusion detection in Ad hoc networks. Clustering Algorithm is used to divide the normal/anomalous data from network and system behaviors. Then HGA is used to dig out the invasion rules. Our EIDA is an unsupervised anomaly detection algorithm. The experiment result shows that it is extensible and not sensitive to the sequence of the input data sets. It has the capacity to deal with different types of data and detection rate and false positive rate of intrusion detection has been improved effectively.

Keywords*: Ad hoc Networks, intrusion detection, clustering algorithm, HGA*

1. Introduction

Ad hoc network is a self-organized network with dynamic topologic and free movement of nodes. The features of Ad hoc network such as centricity and medium openness, unreliable public wireless channel, dynamic changing topology, rely on node routing mechanism of mutual cooperation, no monitoring and management center, etc., made it vulnerable to various attacks. Intrusion detection in the network routing security plays an important role of “the second firewall” [1].

Recently, many researchers propose different intrusion detection models and methods for Ad hoc networks. However there are some main drawbacks of existing intrusion detection model, such as adaptive ability weakness, sensitive to the order of the input data set and can’t detect some new or unknown forms of invasion, detection rate, false positives rate and non-response rates. All of these problems remain to be further improved.

In this paper, we propose a hybrid approach for intrusion detection, which combine improved clustering with HGA, we called enhanced intrusion detection algorithm (EIDA). Firstly, we apply improved cluster algorithm to differentiate normal/abnormal behaviors of network nodes, then HGA is used to dig out normal/abnormal intrusion modes and update rule base of intrusion detection, finally a real-time intrusion detection rule base is set up via the hybrid approach of intrusion detection and EIDA.

The rest of the paper is organized as follows. In section 2, we present the relevant background information and intrusion detection techniques in Ad hoc networks. In section 3, we analyze the security problems in Ad hoc networks in detail. The improved clustering algorithm and HGA for Intrusion Detection is proposed in section 4. Section 5 presents experiment and evaluation results to certify the validity of our algorithm. Finally, section 6 concludes the paper with a brief summary and describes our future research directions.

Received May 28, 2013; Revised July 15, 2013; Accepted August 28, 2013

2. Related Work

Data mining is an efficient method for intrusion detection, which can dig out the unknown knowledge and rules from a large number of network data or audit data from host. Big security data is analysed by some specific data mining algorithm. With the continuous development of data mining technology and highly intelligent itself, it's a good choice to apply data mining technology to Ad hoc network for intrusion detection.

Wenke Lee [2] first adopted data mining technology to intrusion detection. Correlation analysis had been used for intrusion detection to find out some relations among attack attributes. Barbara [3] focused on the research of correlation analysis for intrusion detection and new multiple detection methods had been presented in the system. Mohiuddin et al. [4] presented a program based on Bayesian classification techniques for intrusion detection, which first extracts features from data marked as normal or abnormal, then generates rules for these features, finally compares the tested data from rules to find abnormal or not. Portnoy [5] used clustering techniques to intrusion detection analysis, which method is an unsupervised learning. Bhavani [6] discussed the differences of all kinds of general security threats and analyzed these types of problems with data mining technology. A novel collaborative detection scheme [7] was developed through using data mining anomaly detection technique that enables the IDS to correlate local and global alerts. Gelbard et al. [8] proposed a new hierarchical aggregation algorithm, which the classified data stored in a two-dimensional matrix in the binary form, called binary-positive method.

For the shortcomings of traditional k-means algorithm, a series of improved k-means algorithm are proposed. Ding et al. [9] extended nearest neighbor consistency concept to data clustering, and proposed K-means-CP algorithm. J. Xie et al. [10] combined genetic algorithm with k-means algorithm for the optimization. Phyu [11] proposed a new data sample distribution k-means algorithm to select initial cluster centers. Thomas et al [12] adopted iterative K-Means clustering algorithm together with Map-Reduce to improve the computational efficiency to find the best cluster.

Numerous research works focuse on the security problems in Ad hoc network by using anomaly intrusion detection to prevent attackers from enrolling in the network. An anomaly intrusion detection model based on genetic neural network was presented in [13], which combines the good global searching ability of genetic algorithm with the accurate local searching feature of BP networks for the algorithm optimization. An improved self adaptive Bayesian algorithm (ISABA) was adopted to the alert information classification reducing false positives in intrusion detection [14]. The traditional Boyer-Moore (BM) algorithm was improved in [15] to mobile agent for their collaborative processing intrusion detection. A secure alarm information exchange framework was set up for Ad hoc network routing protocols in [16]. Several anomaly detection approaches for Ad hoc network were evaluated and compared in [17].

Traditional clustering algorithms are sensitive to the input order of intrusion detection data and different initial clustering centers will lead to different clustering results, even falling into local optimal solution. Aim to solve these problem, distance is adopted to describe the differences between the data. We propose an enhanced performance of intrusion detection algorithm, called EIDA. Clustering algorithm is improved firstly, where heuristic clustering method is to solving the clustering number *k*. It's scalable and reduces the requirements for the order of the input data set. Then using HGA to mine intrusion patterns at the same time, update and modify the rules base of intrusion detection to improve the performance of detection.

3. Security of Ad hoc Networks

3.1. Vulnerability Analysis

The features self-organizing and multi-hops of Ad hoc network make each node not only has the ability of general mobile terminal, but also can routing and transfer the packet. Node acts as a router and plays the role of exchange network topology information between neighbors. Via the information advertise, each node can create, delete, and update the network routing table. This characteristic makes Ad hoc network different from other networks, but it's also one of its major defects. An attacked node can send false route information or stop the transmiting routing information. So routing protocol of Ad hoc network is vulnerable.

The vulnerability of routing protocols in Ad hoc Networks is as follows.

(1) Transmit Way. each node in Ad hoc network acts as a router, which has no secure end-to-end routing strategy. Meanwhile each node responsible for forwarding packet. This special structure makes Ad hoc network vulnerable to be attacked.

(2) Dynamic of Network Topology. The moblity of node make network topology to be highly dynamic. It also take more complicated routing security issues, for the reason of the neighbor nodes may not be fully trusted.

(3) Wireless Channel Communication. Ad hoc network is a wireless self-organizing network. It is open for all receiver within the range of wireless signal. Malicious node can easily enroll in the network to intrude the network.

(4) Implicit Trust Relationships of Neighbor Nodes. Ad hoc routing protocols don't take much consideration on the security, in which all the nodes are fully trusted. Intrusion nodes can insert wrong routing updating information or broadcast wrong routing and make the network paralysis.

(5) Nodes Denial of Service. For each of node in Ad hoc network should act on the role of transmit packet, while some of nodes likely being energy insufficiency or attacked lead to not forwarding their received packets on purpose or unavailable. In this case, the network easily encounters denial of service attack.

(6) No Monitor Mechanism for Intrusion Nodes. Routing protocol has not the ability to detect and isolate invasion node in Ad hoc network.

3.2. Security Threat and Attack

The security threats in Ad hoc network include routing attacked, resource exhausted and data flow destroyed attack and so on. The routing information is the main data to maintain Ad hoc network topology, so malicious nodes often launched a variety of routing attacks. Via inserting attcked packet or broadcasting false routing message, all the packets will be transmited to the malicious node which can stop transmiting all or part of these packets to the destination node. These attacks lead to serious network over congestion and channel conflict to decrease network availability.

Some of general attacks on routing protocol is as follows.

(1) Blackhole: Malicious node send the false information of shortest path to the destination node so that the packets flow are continuously transmited to it, which cause information “Blackhole”.

(2) Forge Routing Table: Malicious nodes forge routing information and broadcast it, causing the normal nodes routing error.

(3) Denial of Service: Aim at destroying the network routing to make the network undoing for example routing table overflow attacks.

(4) Wormhole: also known as Tunnel attack, which make a serious attack to Ad hoc network routing through a private channel between two malicious collusion nodes.

(5) Sybil: the malicious nodes pretend several identities to destroy the reliability of the routing protocol at the same time.

4. Hybrid Strategy for Intrusion Detection

We propose a hybrid approach combined improved clustering algorithm with Hybrid Genetic Algorithm (HGA), called Enhanced Intrusion Detection Algorithm (EIDA) for intrusion detection. Clustering Algorithm is used to divide the normal/anomalous data from network and system behaviors. Then HGA is used to dig out the invasion rules.

4.1. Improved Clustering Algorithm

Clustering is a kind of unsupervised learning algorithm, which has no marked training data as classification algorithms need. The *k*-means is well known Clustering algorithm [18]. It is a kind of indirect clustering method based on similarity measure between samples, where *k* is the number of Cluster, then *n* data objects are divided into *k* clusters with minimum standard deviation.

Usually adopt P dimensional vector set $X = \{x_1, x_2, ..., x_n\} \in R_p$ to represent the data set, where, $x_i, 1 \le i \le n$ is a sample or object in the data set. Clustering analysis is to find the vector set $U = \{u_1, u_2, ..., u_n\} \in R_p$. X will be divided into k cluster (X_i is i^{th} cluster). Stochasticly select

k initial solutions, then iteratively moving cluster centers to optimize the sum of squared deviations of each cluster. The sum of squared deviations can be calculated by the equation (1).

$$G(X,U) = \sum_{i=1}^{n}\sum_{j=1}^{k}\left\|x_i - u_j\right\|^2 \quad (1)$$

Intrusion detection data has classific and numeric attribute. The classific attribute values are discrete. Euclidean distance is often used to measure the difference degree of numeric attributes. Ke et al. [19] proposed the difference degree calculation method which calculate classifica and numeric attribute separately and then combining into a sum of distance. A *k*-prototype algorithm difine difference degree as the the sum of different weight of numeric attributes distance and classific attribute distance [20], where difference degree of numeric attribute is caculated by the Euclidean distance and classific attribute distance is defined as the number of attribute unmatched.

In this paper, the traditional clustering algorithm was improved to be able to deal with unlabeled abnormal data samples for intrusion detection. The number of clustering is not preset but determined automatically in the process of clustering. Distance is used to describe the differences between data. The classific and numeric attribute are treated separately, where classific attribute distance is calculated by information entropy and the other by Euclidean distance.

Assumed that the data set X has n objects, each object has m characteristic attributes (include p classific attributes and q numeric attributes), $m = p + q$, then the Difference degree between object x_i and cluster U_j can be calculated by the equation (2).

$$dist(x_i,U_j) = d_p(x_i,U_j) + d_q(x_i,U_j) \quad (2)$$

where $d_p(x_i,U_j)$ is the classific attribute distance between x_i and U_j, $d_p(x_i,U_j) = -\sum_{i=1}^{p} H(a_i)$. $H(a_l)$ is the information entroy of attribute a_l. $d_q(x_i,U_j)$ is the numeric attribute distance between x_i and U_j, $d_q(x_i,U_j) = \sqrt{\sum_{l=1}^{q}(x_{il} - U_{jl})^2}$. Where x_{il} is the l^{th} attribute value of the object x_i, U_{jl} is the l^{th} attribute value of the center of cluster U_j, $1 \le i \le n, 1 \le j \le k$.

For each data object x_i in a given input set $X = \{x_1, x_2, ..., x_n\} \in R_p$, The first step is to find most closed cluster. If the distance is less than the radius threshold R of cluster, then put it in the cluster, else creat a new cluster. Our inproved cluster algorithm is described as Table 1.

The result of clustering algorithm is to set up a number of clusters, containing part of the connection record in each cluster. Due to the difference between normal connection records and abnormal ones, they should be put in a different cluster. Thus we can mark cluster containing abnormal connection records as abnormal, and the other one as normal.

4.2. Create and Update Rules

The cluster centers resulted by improved clustering algorithm can be as the characteristics data set. The initial populations are created according to these characteristics. In this section, we use Hybrid Genetic Algorithm (HGA) to train input data set in order to set up invasion models and update the rules base of intrusion detection. Initial populations need several generations of evolution. The process of each generation evolution is first calculate the fitness function of every rule, then select the most appropriate rules based on fitness measure, crossover operations and finally generate the next populations after mutation. This generational process is repeated until a termination condition has been reached.

Table 1. The improved k-means algorithm

Algorithm: Improved k-means
Input: Data set $X=\{x_1,x_2,...,x_n\}$
Output: Cluster $U=\{u_1,u_2,...,u_k\}$
Initialize $U=\Phi$
Input x_1 , Create a new cluster U_1 of x_1 , Cluster number *k*=1.
if $X!=\Phi$
Input x_i . Calculate $dist(x_i,U_j)$ Find $dist(x_i,U_{min})$
/* for all $U_j \in U$, $j \neq min$, $dist(x_i,U_j) > dist(x_i,U_{min}), 2 \le i \le n, 1 \le j \le k$.*/
if $dist(x_i,U_{min}) > R$
/* R is the cluster radius threshold*/
k=*k*+1, create a new cluster U_k of x_i
else Add x_i to U_{min}
Update U_{min}
/* for x_i , U_{min} is the cluster with minimum value of $dist(x_i,U_j)$ */
else end
Return U

Every intrusion detection rule is the form of “if-then” containing conditions and results. The network feature attribute is connected by a logic symbol "AND", composing the condition of rules. “IsAttack” as a result indicates attacked or not. A rule in this form is given follow, in which character strings (within double quotes) only describe easyly and will be replaced by real data in practical application.

If (Duration = “ANY” and Protocol_type = “TCP” and Service = “telnet” and Logged_in = false and Count= 0 and Num_root = “ANY” and Num_access_files= “ANY” and Srv Count = 0 and Serror rate = “ANY”)

Then (IsAttack = “buffer_overflow”)

In order to make the rules more generality, we use the character “#” as a wildcard in characteristics table. So the example above is expressed by chromosome vectors as {#, l, 12, 0, 0, #, #, 0, #, 12}. Each binary string consisted of encoded binary of the attribute values represents a gene. All of these binary strings are put together as a chromosome expressed an association rule.

Association rule mining is to explore rules that can identify relationships among a set of attributes dataset. The support, the confidence and the interestingness measures of rules generated are normal used to prioritize association rules. In the process of evolution, the highest priority rule will be set the higher fitness value to have more opportunities for survival and cross in the competition. Assume that a rule is expressed as: $if\ X\ then\ Y$, the support, the confidence and interestingness measures of the rule is defined as follow differently.

$$S = support(X \Rightarrow Y) = P(X \cup Y) = \left|X\ and\ Y\right| / X$$
$$C = confidence(X \Rightarrow Y) = P(Y/X) = \left|X\ and\ Y\right| / \left|X\right|$$
$$I = interest(X \Rightarrow Y) = lnP(Y/X) / P(Y/not\ X)$$

Where N is the number of training data, $\left|X\right|$ means the number of training data satisfied the conditions X ; $\left|X\ and\ Y\right|$ means the number of data satisfied the conditions X and results Y in the form of “$if\ X\ then\ Y$ ”. S is support degree of rule, C is confidence degree of rule; I is interest degree of rule.

The Fitness function can be defined as the equation (3).

$$Fitness = aS + bC + cI \quad (3)$$

Where, a, b, c are the weights of three measures of rules ($0 \le a, b, c \le 1$) assigned on the user's preference to optimize the evolution process on fiting individuals.

Genetic operators mainly include selection, crossover and mutation. Tournament selection algorithm is used to select the highest fit individual and inherit to the next generation group. Crossover operation uses the method of binary uniform crossover, which exchanges a pair of individual genes in the same crossover probability to generate two new individuals. Mutation operation is binary variation method to randomly select some individual gene to invert.

5. Evaluation and Analysis

KDD Cup99 [21, 22] dataset for intrusion detection is selected as the experimental sample data. These data sets are collected from MIT LL IDS data set of 1998 by team member of IDS laboratory in Columbia University. They model the network traffic, standardize the expression of characteristics, and provide testing and evaluation data sets for evaluating all kinds of intrusion detection algorithms [23]. It is now well known data set for network security research. Figure 1 shows our experiment process.

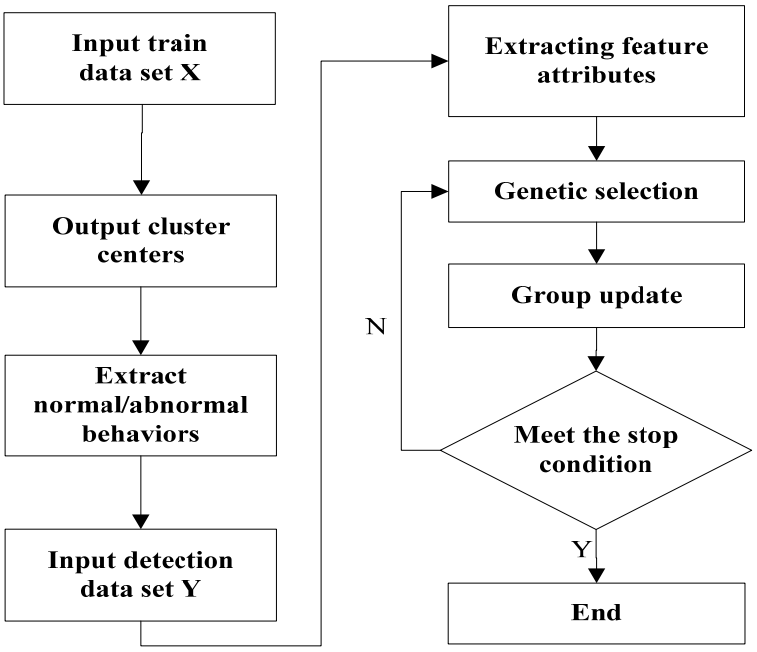

Figure 1. The flow chart of simulation experiment

The original data is too big, so we stochasticly select 7500 records from KDD CUP99 samples as the training set *X*, including 7000 normal data and 500 invasion data. Invasion data includes five types of attacks, such as phf, neptune, land, loadmodule and multihop. Test data set *Y* is randomly selected 12000 data records from the rest of data set, including, containing 10400 normal data and 1600 invasion data with 10 attack types of perl, phf, neptune, teardrop, rootkit, land, loadmodule, multihop, warezmaster and portsweep.

Under the same experimental environment and sample data sets, we test repeatedly six times using stochastic data set to compare our EIDA with other algorithms, GACH in [24], and improved k-means algorithm in [25]. The evaluation results are shown in figure 2 and figure 3.

Figure 2 shows that the detection rate of our hybrid method EIDA is higher than each of other two algoritms. Further more, via several times random tests, we find EIDA is not limited on the sequence of the input data set, to show that it is scalable and our improved k-means algorithm is stable. Figure 3 shows the false detection rate of EIDA is also lower than each of others. And the different initial clustering centers have little impact on the performance of EIDA.

At the same time, the EIDA can dig out normal/abnormal intrusion modes and update rule base of intrusion detection in time.

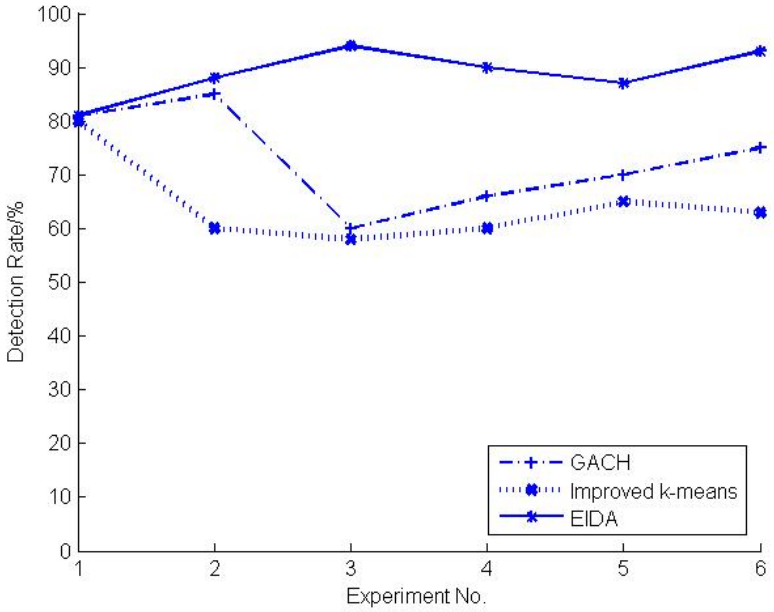

Figure 2. Comparison of detection rate

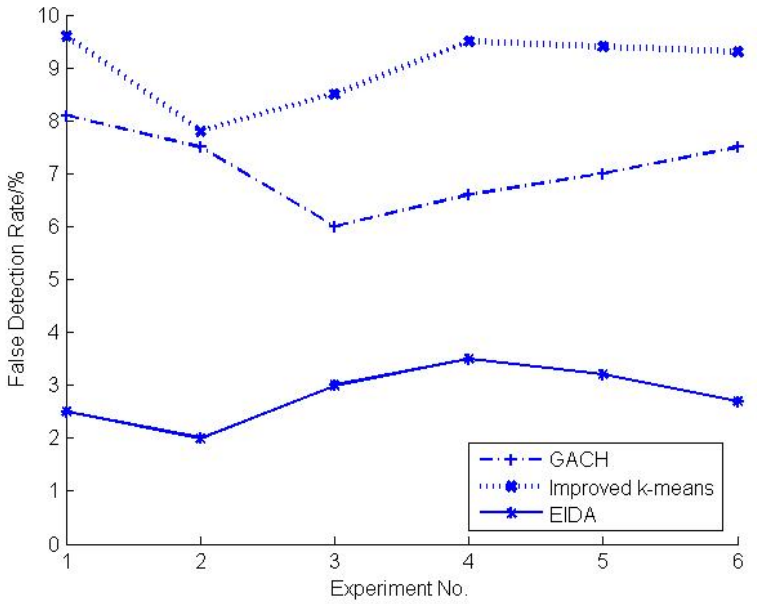

Figure 3. Comparison of false detection rate

6. Conclusion

Ad hoc networks encounter higher safety challenge due to its vulnerability of dynamic topology, limited wireless channel and resources, and so on. For the shortcomings of existing intrusion detection algorithm in Ad hoc network, this paper proposes an hybrid approach for intrusion detection, which combine improved k-means clustering algorithm with HGA, that is an

Enhanced Intrusion Detection Algorithm (EIDA). Firstly, using the improved k-means algorithm to divide the system and network behavior set into normal/abnormal behavior base, then using the HGA to extract invasion modes from the normal/abnormal behavior base, automatically updating intrusion detection rule base. Experiment results show that the algorithm effectively improves the intrusion detection rate and false positive rate. In the future work, we will focuse on how to improve the accuracy of clustering and timely extrac rule, further to increase intrusion detection rate and apply our proposed hybrid intrusion detection algorithm to IDS in Ad hoc network.

Acknowledgment

This work is supported by the State High-Tech Development Plan (No.2013AA01A601).

References

[1] Y Zhang, W Lee. *Intrusion Detection in Wireless Ad Hoc Network*. Proc. MOBICOM 2000, Boston, ACM press. 2000: 275-283.

[2] Wenke Lee, Salvatore J Stolfo. *A Data Mining Framework for Building Intrusion Detection Models*. Proceedings of the 1999 IEEE Symposium on Security and Privacy. 1999: 120-130.

[3] Barbara D. *Detecting intrusions by data mining*. Proceedings of the 2001 IEEE Workshop on information Assurance and security. 2001: 11-16.

[4] Mohiuddin S, Hershkop S, Bhan R, Stofo S. *Defending against a large scale denial of service attack*. Proceedings of the IEEE. Workshop on Information Assurance and Security. 2002: 17-19.

[5] Portnoy L, Eskin E, Stolofo S. *Intrusion detection with unlabled data using clustering*. Proceedings of the ACM CSS Workshop on Data Mining Applied to Security. 2001: 5-8.

[6] Bhavani T. *Data Mining for Malicious Code Detection and Security Applications*. Web Intelligence and Intelligent Agent Technologies. 2009; 31(2): 88-100.

[7] Liu Yu, Li Yang, Man Hong. A distributed cross-layer intrusion detection system for ad hoc networks. *Annales des Telecommunications/Annals of Telecommunications*. 2006; 61(4): 357-378.

[8] Gelbard R, Goldman O, Spiegler I. Investigating diversity of clustering methods: An empirical comparison. *Data & Knowledge engineering*. 2007; 63(1): 155-166.

[9] Ding C, He X. *K-Nearest-Neighbor in data clustering: Incorporating local information into global optimization*. Proc. of the ACM Symp. on Applied Computing. Nicosia: ACM Press. 2004: 584-589.

[10] J Xie, J Wu, Q Qian. *Feature selection algorithm based on association rules mining method*. In Eighth IEEE/ACIS International Conference on Computer and Information Science. 2009: 357-362.

[11] TN Phyu. *Survey of Classification Techniques in Data Mining*. Proceedings of the International MultiConference of Engineers and Computer Scientists. 2009: 727-731.

[12] Thomas L, Manjappa K, Annappa B. *Parallelized K-Means clustering algorithm for self aware mobile Ad-hoc networks.* ACM International Conference on Communication, Computing and Security. 2011: 152-155.

[13] H Jiang, JH Ruan. The Application of Genetic Neural Network in Network Intrusion Detection. *Journal of Computers*. 2009; 4(12): 1223-1230.

[14] Farid DM, Rahman MZ. Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm. *Journal of Computers*. 2010; 5(1): 23-31.

[15] ZS Hou, Z Yu, W Zheng. Research on Distributed Intrusion Detection System Based on Mobile Agent.*Journal of Computers.* 2012; 7(8): 1919-1926.

[16] Shiau-Huey Wang. An Exchange Framework for Intrusion Alarm Reduction in Mobile Ad-hoc Networks. *Journal of Computers*. 2013; 8(7): 1648-1655.

[17] Panos Christoforos, Stavrakakis Ioannis. *An evaluation of anomaly-based intrusion detection engines for mobile ad hoc networks*. 8th International Conference on Trust Privacy and Security in Digital Business, TrustBus. 2011; 6863: 150-160.

[18] Zhang qingwei, Almulla Mohammed. *An efficient certificate revocation validation scheme with k-means clustering for vehicular ad hoc networks*. Proceeding-IEEE Symposium on Computers and Communicatons. 2012: 862-867.

[19] Ke W, Salvatore JS. *Anomalous payload based network intrusion detection*. Proc of the 7th International Symposium on Recent Advanced in Intrusion Detection (RAID). 2004: 201-222.

[20] Lippmann RP, Fried DJ, Graf I. *Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation*. Los Alamitos, CA: Proc of the 2000 DARPA Information Survivability Conference and Exposition (DISCEX). 2000: 12-26.

[21] Engen Vegard, Vincent Jonathan, Phalp Keith. Exploring discrepancies in findings obtained with the KDD Cup99 data set. *Intelligent Data Analysis*. 2011; 15(2): 251-276.

[22] Chandrashekhar AM, Raghuveer K. Performance evaluation of data clustering techniques using KDD Cup99 Intrusion detection data set. *International Journal of Information and Network Security (IJINS)*. 2012; 1(4): 294-305.

[23] LI Xiangyang, YE Nong. A supervised clustering and classification algorithm for mining data with mixed variables. IEEE *Trans on Systems, Man and Cybernetics*. 2006; 36(2): 396-406.
[24] SUN B, OSBORNE L, YANG Xiao. *Intrusion detection techniques in mobile Ad Hoc and wireless networks*. IEEE Wireless Communications. Piscataway IEEE. 2007: 56-63.
[25] Jian Li, Jun Li. *Application Research of Improved K-means Clustering Algorithm in Intrusion Detection System*. 2011 National Communications Security Conference of China. 2011: 127-131.