International Journal of Security and Its Applications
Vol.9, No.11 (2015), pp.239-250
http://dx.doi.org/10.14257/ijsia.2015.9.11.23

Network Intrusion Detection Model With Clustering Ensemble Method

Liang-Wei Chen

Department of information and Engineering, Chengdu Aeronautic Polytechnic, Chengdu, Sichuan, China, 610100
clw206064@163.com

Abstract

As network techniques have become one of the most significant aspects of our daily lives, network security has been a major concern. One common application is network intrusion detection. From the perspective of data oriented consideration, intrusion detection can be formulated as a clustering task, which aims to differentiate normal and insecurity behaviors and categorize into several groups. In this paper, we employ ensemble clustering method to improve the generalization and robustness of basic clustering. Specifically, we employ fuzzy kernel C-means (FKCM) as basic clustering, which improves the fuzzy C-means (FCM) clustering by introducing kernels from the support vector machines (SVM) to optimize the features of sample data by mapping the sample pattern into a higher dimensional feature space. Then, we formulate the ensemble problem as the optimization of the mutual information among all clusterings and introduce Ant Colony Optimization (ACO) as the solution. Experiments prove the efficiency of our method.

Keywords*: Network intrusion detection, Ensemble clustering, Fuzzy kernel C-means (FKCM), Ant Colony Optimization (ACO)*

1. Introduction

With the rapid development of communication and Internet techniques, networks have been one of the most important tools in everyday life. However, it is also a major concern for governments, enterprises and network users that network can be attacked and network information can be leaked, modified and misused for unsafe and illegal purposes [1].

One of the most interesting efforts on network security is network intrusion detection [2], which refers to actively collect and analyze network data, and then detect behaviors or signs of network attacks and other forms of insecurity. From the perspective of data oriented applications, network intrusion detection is a process of data analysis. Specifically, in this work, we formulate intrusion detection as a clustering task, which aims to differentiate normal and insecurity behaviors and categorize into several groups.

As one of the most important data mining methods, fuzzy clustering techniques have been widely applied in many fields, such as data analysis [3], pattern recognition [4] and image processing [5], *etc*. Specifically, in this paper, we employ fuzzy kernel C-means (FKCM) clustering [6], which improves the fuzzy C-means (FCM) [7] clustering by introducing kernels from the support vector machines (SVM) to optimize the features of sample data by mapping the sample pattern into a higher dimensional feature space. Indeed, FKCM can magnify the differences between samples, and therefore improve the clustering accuracy and convergence speed.

Ensemble learning [8] integrates multiple models and techniques to solve one single problem, and can significantly improve the generalization and robustness of the learning process. In this paper, after generating clustering results from the basic clustering

ISSN: 1738-9976 IJSIA

algorithm on multiple training samples, we formulate the problem as the optimization of the mutual information among all clusters and introduce Ant Colony Optimization (ACO) [9] as the solution.

2. Related Work

The first category of related work is network intrusion. Intrusion Detection System (IDS) was first proposed by [10], and later Gates *et a*l. [11] constructed an IDS model. Mathew *et a*l. [12] Designed an anomaly detection method based on correlation rules. Cucurull *et a*l. [13] summarized the comparison of multiple intrusion detection techniques on SWITCH/AS559 dataset. Nehinbe *et a*l. [14] proposed Expert-track model to automatically adjust the rule space. Caberera *et a*l. [15] constructed a detection method based on Poisson distribution to detect DoS attacks. However, the accuracy still needs improving. In this paper, we employ clustering ensemble with ACO as the ensemble strategy for network intrusion detection.

The second category of related work is ensemble clustering. The basic strategies of ensemble learning include bagging, boosting and AdaBoost, etc. Applying ensemble learning in the field of clustering is called ensemble clustering. The efforts on ensemble clustering lie in the combination of basic clusterings. For example, Fred *et a*l. [16] proposed to use co-association matrix for combining clustering results. Tang *et a*l. [17] used a weighted selective voting strategy for combination. Strehl *et a*l. [18] introduced graph theory method. Ayad *et a*l. [19] proposed a probability model based on metric in information theory, *i.e.*, Jensen-Shannon Divergence. In this work, we formulate the ensemble process as an optimization problem and introduce intelligence algorithm ACO as a solution.

3. Preliminaries

FCM clustering is one of the common clustering techniques, which uses membership degree to determine if an element belongs to a cluster. Let $\{x_i, i = 1,2,...,n\}$ be the set of sample data, where n is the size of sample, and $\{o_j, j = 1,2,...,c\}$ be the centroids of clusters, where c is the number of clusters. The objective function of FCM clustering is:

$$J = \sum_{j=1}^{c}\sum_{i=1}^{n}(u_{ij})^2 \left\| x_i - o_j \right\|^2 \quad , \tag{1}$$

where u_{ij} is the membership degree of data i to cluster j, and

$$\sum_{j=1}^{c} u_{ij} = 1, \forall i \in [1,n] \quad . \tag{2}$$

The goal of FCM is find the optimal clustering results which minimize the objective function, and the process of FCM can be summarized as follows:

Step 1: initialize the centroids of clusters as $O = \{o_1, o_2, ..., o_c\}$.

Step 2: calculate the membership degree matrix:

$$u_{ij} = \left[\sum_{k=1}^{c} \left(\frac{d_{ij}(x_i - o_j)}{d_{ik}(x_i - o_k)} \right)^{\frac{2}{r-1}} \right]^{-1}, k = 1,2,...,n \ , \quad (3)$$

where r is the coefficient of weights.

Step 3: update the centroids as:

$$o_j = \frac{\sum_{i=1}^{n} u_{ij}^{'} x_i}{\sum_{i=1}^{n} u_{ij}^{r}} \quad . \quad (4)$$

Step 4: repeat Steps 2 and 3 until the differences of membership degree matrix between iterations is smaller than given ε.

Kernel based method is a nonlinear method for data processing, which uses a non-linear function to map the original data into a higher dimensional space. Let the original data sample $x_i \in R_p (i = 1,2,..., p)$ be mapped to feature space H by a kernel function ϕ, that is, $\{\phi(x_1), \phi(x_2), ..., \phi(x_p)\}$, and the feature space can be represented as:

$$K(x_i, x_j) = \left\langle \phi(x_i) \cdot \phi(x_j) \right\rangle \quad , \quad (5)$$

where $K(x_i, x_j)$ is the kernel function.

4. Proposed Intrusion Detection Model

Generally, we employ clustering ensemble method for intrusion detection, which refers to first generate a set of clustering results using a basic clustering algorithm, and then combine those results together using a specific ensemble method.

Generally, there are two main tasks in ensemble learning: (1) generating basic clustering results and (2) ensembling multiple basic clusterings. In this paper, we use FKCM clustering as the basic clustering algorithm. Then, we introduce the idea from ACO as the ensemble strategy with the objective of minimizing the distances between clusters.

4.1 Basic Clustering Using FKCM

In his paper, we employ FKCM algorithm for basic clustering, since FKCM improves FCM by introducing a kernel function. Basically, the major difference is a kernel function matrix composed by data sample pairs as Equation (5).

Given data sample $X = \{x_i, i = 1,2,...,n\} \in R_d$, where n is the size of samples, d is the dimensions of sample data. Suppose c is the number of clusters, and $V = \{v_1, v_2, ..., v_c\}$ is the centroids of each cluster, $U = (\mu_{ik})_{c \times n}$ is the fuzzy

membership degree matrix, and element μ_{ik} is the membership degree of sample k to cluster i.

The clustering process can be summarized as an optimization problem of clustering criterion function, and the objective is to learn the centroid matrix V and membership degree matrix U given data sample X. Therefore, the objective function is:

$$\begin{aligned} J(X,U,V) &= \sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}^{m}\left\|\phi(x_k)-\phi(v_i)\right\|^2 \\ &= \sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}^{m}[K(x_k,x_k)-2K(x_k,v_i)+K(v_i,v_i)] \end{aligned}, \quad (6)$$

where $K(x_k,x_k)$ is the kernel function value of sample k, $K(v_i,v_i)$ is the kernel function value of centroid i, $K(x_k,v_i)$ is the kernel function value of sample k and centroid i, m is weighted index number, and typically $m=2$.

In order to minimize Equation (6), using Lagrange multipliers, we iterate membership degree with the following equation:

$$\mu_{ik}=\sum_{j=1}^{c}\left[\frac{K(x_k,x_k)-K(x_k,v_i)+K(v_i,v_i)}{K(x_k,x_k)-2K(x_k,v_j)+K(v_j,v_j)}\right]^{\frac{2}{m-1}}, \quad (7)$$

where $K(x_k,v_i)$ is the kernel function value of sample k and any cluster j, and $K(v_j,v_j)$ is the kernel function value of any cluster j.

And the centroid of cluster j is updated as:

$$V_i=\frac{\sum_{k=1}^{n}\mu_{ik}^{m}K(x_k,v_i)x_k}{\sum_{k=1}^{n}\mu_{ik}^{m}K(x_k,v_i)}. \quad (8)$$

In short, as shown in Figure 1, the process of FKCM can be summarized as follows:

Step 1: given the number of clusters c, and initialize kernel function parameters. Set the threshold of algorithm termination as ε, and the iteration count as $t=0$;

Step 2: update the fuzzy membership degree matrix $U^{(t)}$;

Step 3: update the centroid matrix $V^{(t)}$ as Equation (8);

Step 4: if $\|V^{(t+1)}-V^{(t)}\|<\varepsilon$, algorithm stops and output U,V; otherwise, set $t=t+1$, and return to Step 2.

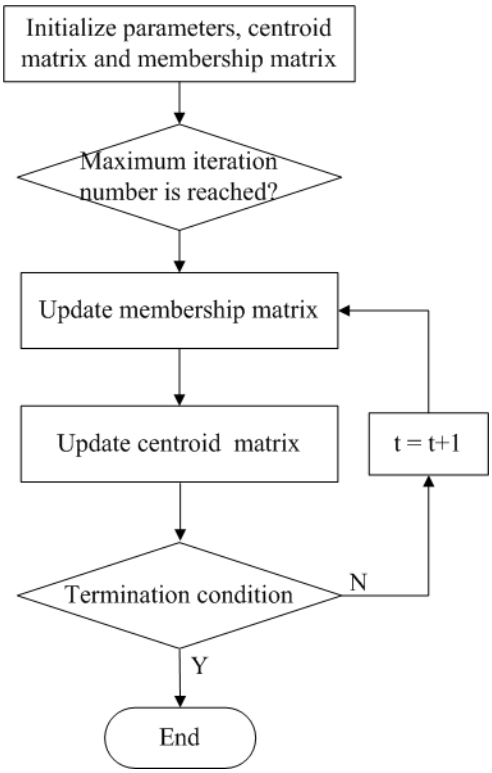

Figure1. Process flow of FKCM

4.2 Clustering Ensemble with ACO

The idea of clustering ensemble can be described as follows. Suppose the dataset $D = \{x_1, x_2, ..., x_n\}$.

Step 1: apply basic clustering algorithm on D N times, and thus get N clustering results $P = \{P_1, P_2, ..., P_N\}$, where P_i is the clustering result of i-th clustering.

Step 2: combine the results of P as a new partition solution P_o using some strategy.

Step 3: for each data sample, determine which cluster it belongs to based on optimal clustering P_o.

In this paper, we use bagging [20] technique for generating basic clusterings. That is, randomly select multiple samples from the dataset, so that all sample datasets are of similar size. Note that data points can be repeatedly selected into multiple training samples; that is, a data point could be in multiple training samples, or not in any samples at all. In order to generate multiple clusterings with more difference, we first use bagging to generate multiple training samples, and then FKCM clustering is performed on each training sample.

In order to ensemble all clustering results from basic FKCM clustering, we construct a criteria function based on the distances of clustering members, and then employ ACO algorithm to find the optimal solution which minimize the distance.

ACO was first proposed by Dorigo [9], which is a heuristic optimization technique with positive feedback mechanism, inspired by the food searching behavior of ant colony. The basic idea can be described as follows: when an ant passes a route, it leaves pheromones on it, and the more pheromones the route has, the more likely it would be chosen by other ants. In other words, the probability of routes being selected is related to the amount of pheromones remained, which contributes to positive feedback in return. That is, ants communicate with each other by releasing pheromones on routes when searching for food.

Suppose the phenomenon of ant k on the route from node i to j at time t is notated as $\tau_{ij}^{k}(t)$, and the initial phenomenon on route (i, j) is $\tau_{ij}^{k}(0) = c$, where c is a constant. Let the probability of ant k transferring from node i to j at time t be $p_{ij}^{k}(t)$, calculated as:

$$p_{ij}^{k}(t) = \begin{cases} \dfrac{\tau_{ij}^{k}(t)^{\alpha}\left(1/d_{ij}^{k}\right)^{\beta}}{\sum_{s \in allowed_k} \tau_{ij}^{s}(t)^{\alpha}\left(1/d_{ij}^{s}\right)^{\beta}}, & j \in allowed_k; \\ 0, & otherwise. \end{cases} \quad (9)$$

where $allowed_k$ is the set of nodes that haven't been passed through yet by ant k, α is the heuristic factor, notating the importance of route with remaining pheromones, β is the heuristic factor for $1/d_{ij}^{k}$, notating the affect of heuristic information, and d_{ij}^{k} is the distance of route (i, j) passed by ant k.

At the next iteration $t+1$, the phenomenon on route (i, j) is updated by:

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \Delta\tau_{ij}(t) \quad , \quad (10)$$

$$\tau_{ij}(t) = \sum_{h=1}^{m} \Delta\tau_{ij}^{k}(t) \quad , \quad (11)$$

where $1-\rho$ is the residual coefficient of pheromones, $\rho \subset [0,1)$, and m is the number of ants. $\Delta\tau_{ij}^{k}(t)$ is the amount of pheromones remaining on route (i, j) at current iteration for ant k, which is calculated as:

$$\Delta\tau_{ij}^{k}(t)=\begin{cases}\dfrac{Q}{L_k}, & \text{if ant } k \text{ passes } (i,j) \text{ at current iteration;}\\ 0, & \text{otherwise.}\end{cases} \tag{12}$$

where Q is a constant, and L_k is the total length of ant k 's tour.

Basically, the ensemble method is to define a function based on the mutual information among all clustering results, and then apply ACO to minimize the funciton. We define the distance between clusterings using mutual information to calculate the relevance of two clustering solutions. Given two clusterings $P_{(1)}, P_{(2)}$, define the mutual information between them as:

$$I(P_{(1)},P_{(2)})=\sum_{i,j}p_{ij}\log\frac{p_{ij}}{p_i p_j} \quad , \tag{13}$$

where i, j are the centroids of clusters in $P_{(1)}, P_{(2)}$, p_i, p_j are the marginal probability distribution of $P_{(1)}, P_{(2)}$, and p_{ij} is the joint probability distribution of $P_{(1)}, P_{(2)}$. The larger $I(P_{(1)}, P_{(2)})$ is, the more similar two clusterings are, and therefore the distance between them is smaller.

Since $I(P_{(1)}, P_{(2)})$ is sensitive to the overlap between clusterings, we use Normalized Mutual Information (NMI) instead:

$$NMI(P_{(1)},P_{(2)})=-\frac{\sum_i p_i\log p_i+\sum_j p_j\log p_j}{\sum_{i,j}p_{ij}\log p_{ij}} \tag{14}$$

Therefore, given N clustering results $P=\{P_1,P_2,...,P_N\}$, our objective is to find the best clustering P^*, so that:

$$f(P^*)=-\sum_{i=1}^{N}NMI(P^*,P_i) \quad . \tag{15}$$

The objective is to minimize above equation, and we apply ACO to achieve that goal. Assign ants onto N clusterings, $\tau_{ij}^{k}(t)$ is the pheromone on the route from clustering solution i to j for ant k, and $\eta_{ij}^{k}(t)$ is the expectation of route (i, j).

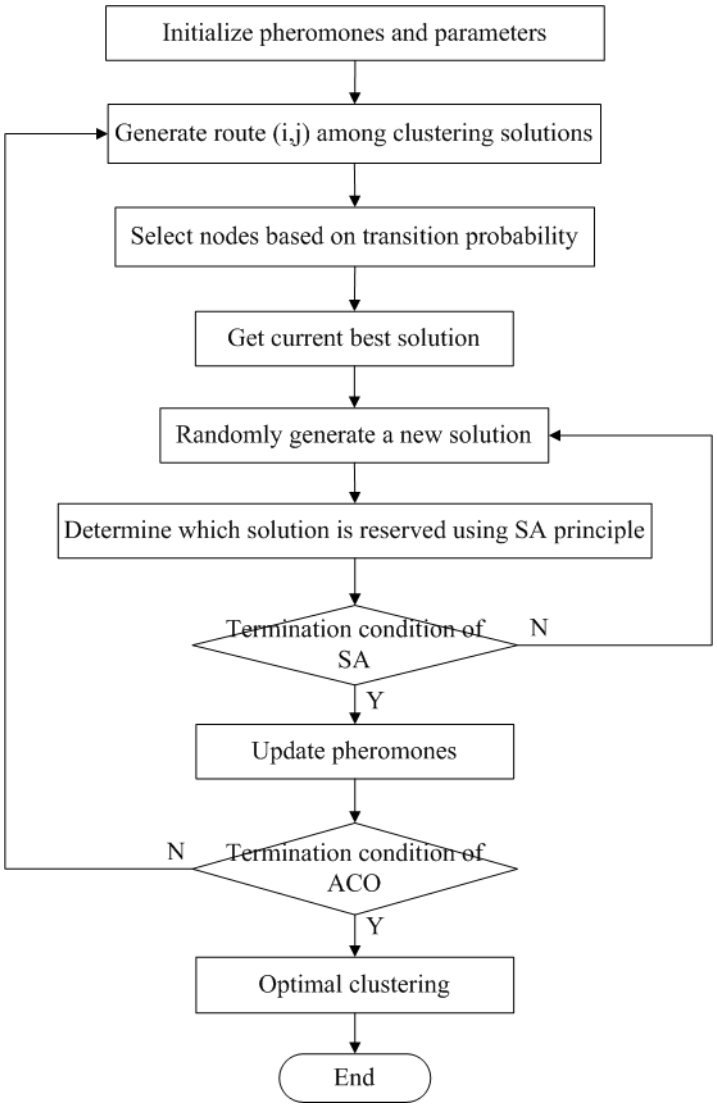

Figure 2. Process Flow of ACO Based Ensemble Clustering

However, typical ACO is prone to arriving at local optimal point, and the convergence speed is typically low. Local best solution of node selection sequence X for ants is obtained after each iteration of node selection, typical ACO use this sequence for pheromone updates directly until convergence or the maximum iteration number is satisfied. In this work, we introduce the idea of Simulated Annealing (SA) [21] to avoid local optimal and accelerate the convergence speed. Specifically, we introduce the SA idea for pheromone updates.

As shown in Figure 2, the process of improved ACO is described as follows.

Step 1: randomly exchange two clustering results to generate a new solution X'.

Step 2: calculate fitness values for X, X'. If $F(X)-F(X')<0$, replace X with X'; otherwise, if $\exp\left(-(F(X)-F(X'))/T\right) > rand(0,1)$, replace X with X'; otherwise, preserve X.

Step 3: $T(t+1) = L \cdot T(t)$, where T is the annealing temperature, and L is the annealing coefficient.

Step 4: if $T < T_{end}$, where T_{end} is the stopping temperature, go to Step 5; otherwise, go to Step 1.

Step 5: update pheromone as Equations (10) and (11).

5. Experiment

We use KDD CUP99 [22] dataset for evaluation, which includes four kinds of network intrusions: Probing, R2L, U2R and DoS. In our experiment, we sample a subset of KDD CUP99 dataset. We extract 10,000 records for training set, which includes 100 intrusion records. As for test set, we select two different test datasets: (1) a small sample chosen from training set, (2) unseen intrusion data from the remaining data. That is, we test the performance when the intrusion is known and unknown respectively.

There are 41 attributes in the original dataset. As indicated in [23], only 13 attributes are significant, including 11 numerical attributes and 2 categorical attributes. Since FCM algorithm cannot deal with categorical attributes directly, we first transform them into numerical attributes. That is, encode each category of the attribute as one bit. If the attribute value equals to i-th category, set the i-th bit as 1, and other remaining bits are 0. For example, attribute protocol_type has three categories: tcp, udp and icmp. Thus, the encoding is 001, 010 and 100 respectively.

Next, we perform data normalization on the dataset. Transform each data as:

$$x'_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j}, \overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2} \quad ,(16)$$

where x_{ij} is the j-th attribute of the i-th data. After the normalization, all values are transformed within the range [0,1].

We use two metrics to measure the performance of detection. First, we define intrusion detection rate (DR) as:

$$DR = \frac{num_{ind}}{num_{in}} \times 100\% \quad , \quad (17)$$

where num_{in} is the number of intrusions in the test set, and num_{ind} is the number of intrusions detected in the test set.

Second, we define error rate (ER) as:

$$ER = \frac{num_{ed}}{num_{norm}} \times 100\% \quad , \quad (18)$$

where num_{ed} is the number of error detected records, and num_{norm} is the number of normal records totally.

We compare the performance of proposed algorithm with different ratios of the number of anomaly clusters to the number of all clusters, named anomaly cluster ratio for short.

Table 1 Comparison with Different Anomaly Cluster Ratios

Anomaly cluster ratio	Test set 1		Test set 2	
	DR	ER	DR	ER
1/6	0.8854	0.0132	0.8162	0.0144
1/5	0.8935	0.0188	0.8329	0.0197
1/4	0.9096	0.0203	0.8472	0.0215

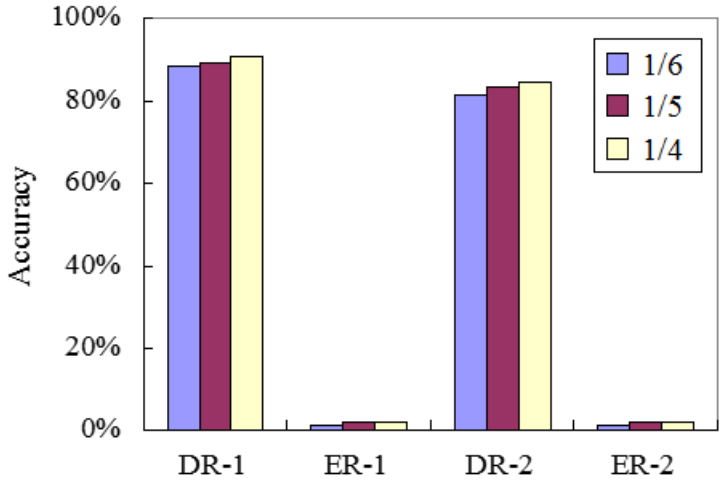

Figure 3. Performance Comparison

We can observe from Table 1 and Figure 3 that when the anomaly cluster ratio is 1/5 the overall performance of intrusion detection is best. Besides, the performance on test data 1 is better than test data 2, since test data 1 is testing known intrusions while test data 2 is testing unknown intrusions. Moreover, with the anomaly cluster ratio increasing, the detection rate and error rate grow. That is, more records are labeled as anomalies, and therefore more intrusions are detected, which also contributes to higher error detection rate.

6. Conclusion

In this paper, we investigate the use of ensemble clustering in the field of network intrusion, and provide the probability of introducing intelligence algorithms as the ensemble strategy. In future, we would like to explore more applications in network security and try to embrace more techniques for solution.

Acknowledgment

Sichuan Province Natural Science Key Project: Application Research on Informationization Construction of the cloud computing, leading researcher of the project: LIANG-WEI CHEN, Project No. 15ZA0399

References

[1] S. Mandala, "Integrating security services into active network[J]", Active network; security services; secure active network; network transport system, **(2014)**.
[2] C. Chen, Y. Chen, H. Lin. "An efficient network intrusion detection [J]." Computer Communications, vol. 33, no. 4, pp. 477 – 484.
[3] C. Döring C, M. Lesot , Kruse R. "Data analysis with fuzzy clustering methods[J"], Computational Statistics & Data Analysis, **(2006)**, pp. 51:192 – 214.
[4] PC B, AB. O. Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data.[J]. Briefings in Bioinformatics, vol. 6, no. 4,**(2005)** pp. 331-343.
[5] M R Rezaee, van der Zwet P M J, Lelieveldt B P E, *et al*. "A multiresolution image segmentation technique based on pyramidal segmentation and fuzzy clustering.[J]", IEEE Trans Image Process, vol. 9, no. 7, **(2000)**, pp. 1238 - 1248.
[6] J, i, a, *et a*l. Fuzzy C-Means Clustering Algorithm Based on Kernel Method[J]. ICCIMA, 2003.
[7] J C, Bezdek, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm[J]. Computers & Geosciences, vol. 10, no. 84, **(1984)**, pp. 191-203.
[8] H . Lappalainen, J W Miskin, "Ensemble Learning [J]", Perspectives in Neural Computing, **(2000)**, pp. 75-92.
[9] M. Dorigo , M. Birattari . "Ant colony optimization[J]. Computational Intelligence Magazine" , IEEE, vol 1, no. 4,**(2006)**, pp. 28 - 39.
[10] T. F. Lunt. "A survey of intrusion detection techniques[J]", Computers & Security, vol. 12, no. 4, **(1993)**, pp. 405-418.
[11] C. Gates , C. Taylor . "Challenging the anomaly detection paradigm: a provocative discussion[C]"//Proceedings of the 2006 workshop on New security paradigms. ACM, **(2006)**: pp. 21-29.
[12] S. Mathew, M. Petropoulos , Ngo H Q, *et a*l. A data-centric approach to insider attack detection in database systems[C]//Recent Advances in Intrusion Detection. Springer Berlin Heidelberg, **(2010)**, pp. 382-401.
[13] Cucurull J, Asplund M, Nadjm-Tehrani S. Anomaly detection and mitigation for disaster area networks[C]//Recent Advances in Intrusion Detection. Springer Berlin Heidelberg, **(2010),** pp. 339-359.
[14] Nehinbe J O. Automated technique for debugging network intrusion detection systems[C]//Intelligent Systems, Modelling and Simulation (ISMS), 2010 International Conference on. IEEE, **(2010)**, pp. 362-367.
[15] J B D Caberera, B. Ravichandran, Mehra R K. "Statistical traffic modeling for network intrusion detection[C]", Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 2000. Proceedings. 8th International Symposium on IEEE, **(2000)**, pp. 466-473.
[16] A L Fred. "Finding Consistent Clustersin Data Partitions[J]"., Multiple Classifier Systems, 2001, 2096: pp. 309 – 318.
[17] T. Wei, Z. Zhihua. "Selective Clustering Ensembling based on Bagging. Software Journal" vol. 16, no. 4**, (2005)** , pp. 496-502. (in Chinese)
[18] A. Strehl , J. Ghosh . Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions[J]. Journal of Machine Learning Research, 2002, 3: pp. 583 – 617.
[19] Ayad H, Basir O A, Kamel M. A probabilistic model using information theoretic measures for cluster ensembles[A]. In: Proceedings of the 5th International Workshop on Multiple Classifier System[C]. Springer, 2004: pp. 144 – 153.
[20] L. Breiman . Bagging predictors[J]. Machine Learning, vol. 24, no. 2, **(1996)**, pp. 123 – 140.
[21] Laarhoven P V, M V L P J, Van Laarhoven P J M, *et a*l. Simulated annealing: theory and applications.[J]. D. Reidel Publishing Co., Dordrecht, **(1987).**
[22] MIT Lincoln Lab.. KDDCUP99 Dataset. http://kdd.ics.uci.edu/databases/kddcup99.
[23] Mukkamala S, Janoski G, Sung A H. Intrusion Detection Using Neural Networks and Support Vector[C]//Proc. of IEEE Int'l Joint Conference on Neural Networks. Honolulu, Hawaii, USA, **(2002).**

