# Intrusion Detection in KDD99 Dataset using SVM-PSO and Feature Reduction with Information Gain

Harshit Saxena
M.Tech. Scholar
Computer Science Engineering
LNCT Bhopal, India

Vineet Richaariya, Ph.D
Computer Science Engineering
LNCT Bhopal, India

## ABSTRACT
Intrusion detection is a process of identifying the Attacks in the networks. The main aim of IDS is to identify the Normal and Intrusive activities. In recent years, many researchers are using data mining techniques for building IDS. Due to the non-linearity and quantitative or qualitative network data traffic IDS is complicated. For making the IDS efficient we have to choose the key features. Support Vector Machine (SVM) gives the potential solution for IDS problem. SVM suffers by selecting the suitable SVM parameters. Here we propose a new approach using data mining technique such as SVM and Particle swarm optimization for attaining higher detection rate. PSO is an Optimization method and has a strong global search capability. The SVM-PSO Method is applied to KDD Cup 99 dataset. Free parameters are obtained by standard PSO for support vector machine and the binary PSO is used to obtain the best possible feature subset at building intrusion detection system. The propose technique has major steps: Preprocessing, Feature Reduction using Information Gain, Training using SVM-PSO. Then based on the subsequent training subsets a vector for SVM classification is formed and in the end, classification using PSO is performed to detect Intrusion has happened or not. The experimental result shows that SVM-PSO acquire high detection rate than regular SVM Method algorithm.

## Keywords
Intrusion detection system; Information Gain; Support Vector Machine (SVM); Particle Swarm Optimization (PSO)

## 1. INTRODUCTION
Conventional Intrusion avoidance technique such as firewall, access control and encryption has failed to detect the intrusion in the networks. As a result Intrusion detection system becomes an essential component. The idea of the Intrusion detection system (IDS) is to prevent the computer system from attack. The IDS is the most essential part of the security infrastructure for the networks connected to the internet because various ways to compromise the stability and security of network. IDS can be classified into two types: Anomaly and Misuse detection. Anomaly detection system creates a database of normal behavior and any deviations from the normal behavior are occurred an alert is triggered regarding the occurrence of intrusions. Misuse Detection system stores the Predefined attack patterns in the database if a similar data and if similar situations occur it is classified as attack. Based on the source of data the intrusion detection system are classified to Host based IDS and Network based IDS. In network based IDS the individual packet flowing through the network are analyzed. The host based IDS analyzes the activities on the single computer or host. The main disadvantage of the misuse detection (signature detection) method is that it cannot detect novel attacks and variation of known attacks. To avoid these drawbacks we go for anomaly based detection methods. With this approach, known and novel attacks can be detected. The problem is that it will generate more false alarms [1]. The intrusion detection method based on unsupervised learning has a high detection rate but also a high False positive rate. Intrusion detection functions include [2]:-
Monitoring and examining both user and system activities.

- Analyzing system configurations and weaknesses.
- Assessing system and file integrity.
- Skill to identify patterns typical of attacks.
- Analysis of irregular activity patterns.
- Tracking user policy violations.

The remaining part of this paper is organized as follows: Section 2Describes IDS in general. Section 3 presents an overview offrequentlyoccurring network attacks, and section 4 discusses related research done so far. Section 5 describes our proposed Method. Section 6 describes the Experimental Setup. Section 7 describe the Conclusion. Section 8 describe the References.

## 2. INTRUSION DETECTION SYSTEM
IDS are system software that detects attack on a network or computer system. IDS are normally classified as Misuse detection and Anomaly detection [3]. In Misuse system the signature of known attacks are stored in database. Any data similar to that data is classified as attacks. Anomaly detection refers to statistical knowledge about normal activity. The anomaly detection approach can be categorized into semi-supervised and unsupervised anomaly detection [4]. Semi-supervised anomaly detection approaches need a set of purely normal training data from which they found the profile of normal behavior. If the training data contains some attacks hidden within it, the approach may not detect future instances of these attacks. On the other hand, unsupervised anomaly detection approaches set up the profile of normal behavior with unlabeled training data that consists of both normal as well as anomalous samples. Intrusions correspond to deviations from the normal activity of system. The anomaly detection system has high false positive/ negative alarm rate compared to misuse detection systems.
Many draw back has in conventional Approach:

- Signature-based IDSs must be automatic to detect each attack and thus must be continually updated with signatures of new attacks.
- Many signature-based IDSs have hardly defined signatures that prevent them from detecting variant of common attacks.
- Anomaly detection approaches usually create a large number of false alarms due to the random nature of users and networks.

- Anomaly detection approaches often need wide "training sets" of system occurrence records in order to characterize normal behavior patterns.
- Application-based IDSs may be weaker than host-based IDSs to being attacked and disabled since they run as an application on the host they are monitoring.

## 3. NETWORKING ATTACK

Attacks were classified, according to the goal of attacker. Each attack type falls into one of the following four categories [5]:

**Denials-of Service (DoS)**denial of service attack is a class of attacks in which an attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine.

**Probing** or Probing is a class of attacks in which an attacker scans a network of computers to gather information or find known vulnerabilities. An attacker with a map of machines and services that are available on a network can use this information to look for exploits.

**User-to-Root (U2R)** User to root exploits are a class of attacks in which an attacker starts out with access to a normal user account on the system and is able to exploit vulnerability to gain Root access to the system.

**Remote-to-Local(R2L)** A remote to user attack is a class of attacks in which an attacker sends packets to a machine over a network-but who does not have an account on that machine; exploits some vulnerability to gain local access as a user of that machine.

**TABLE 1. Identify the Type of Intrusion Detection Experimental Data Of Kddcup99**

| Identify the Type | Meaning | Specific Classification Identification |
|---|---|---|
| Normal | Normal record | Normal |
| DOS | Denial of service attacks | Neptune,pod,land, back,smurf, teardrop |
| Probing | Monitoring and other exploration activities | Ipsweep,nmap,portsweep ,satan etc. |
| R2L | Unauthorized access from remote machine | Imap,ftp_write, Warezclient,multihop, phf,spy,guess_passwd, warezmaster |
| U2R | Unauthorized access to local super user privileges by ordinary users | Loadmodule, buffer_overflow,rootkit, perl |

## 4. RELATED RESEARCH WORK ON IDS

In the last decade various approaches have been developed in order to detect the Intrusion. Earlier there are two approaches, rule based expert system and statistical approaches. A rule based expert system can select well known intrusion with high detection rate but it is difficult to detect new Intrusion and its signature database need to be updated manually and frequently. Statistical based IDS employ various statistical method including Principal component analysis, Cluster and Multivariate Analysis, Bayesian Analysis etc.

To overcome the drawback of rule based expert system and statistical approaches, a number of data mining technique have been introduced. ANN (Artificial Neural Network) is one the most widely used and work successfully on Intrusion detection. Different types of ANN are used in IDS like Supervised, Unsupervised and Hybrid ANN [6].

Jirapummin et al [7] proposed employing a hybrid ANN for both visualizing intrusions using Kohenen's SOM and classifying intrusions using resilient propagation neural networks. Horeis [8] used a combination of SOM and radial basis function (RBF) networks. The system offers generally better results than IDS based on RBF networks alone. Han and Cho [9] proposed an intrusion detection technique based on evolutionary neural networks in order to determine the structure and weights of the call sequences. Chen, Abraham, and Yang [10] proposed hybrid flexible neural-tree –based IDS based on flexible neural tree, evolutionary algorithm and particle swarm optimization (PSO). Empirical results indicated that the proposed method is efficient. For ANN based intrusion detection, hybrid ANN has been the trend. But, different ways to construct hybrid ANN will highly influence the performance of intrusion detection. In [11], Axellson wrote a well-known paper that uses the Bayesian rule of conditional probability to point out that implication of the base-rate fallacy for intrusion detection. In [12], a behavior model is introduced that uses Bayesian techniques to obtain model parameters with maximal a posteriori probabilities.

Following this torrent, we propose an approach for intrusion detection, which is a combination of Information Gain, PSO and SVM techniques to enhance detection precision.

## 5. PROPOSED METHOD

In this section, we elaborate our new proposed approach. Our approach shown in figure 1.

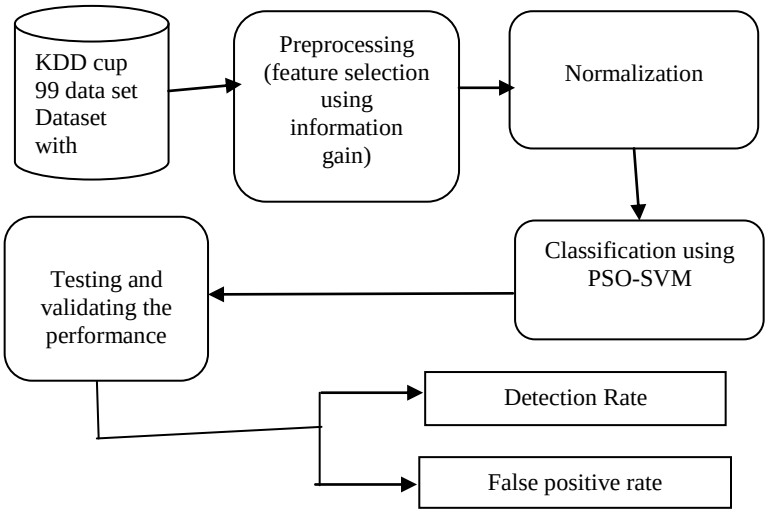


**Figure 1. Block Diagram of our proposed technique**

### 5.1 KDD cup 1999 Dataset

The data set provided for the 1999 KDD Cup was originally prepared by MIT Lincoln labs for the 1998 Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Evaluation Program, with the objective of evaluating research in intrusion detection, and it has become a benchmark data set for the evaluation of IDSs.It Contains approximately 49, 00,000

data instances. This dataset contains 4 900 000 replicated attack on record, there is one type of the normal type of the identity of normal and 22 kinds of attack types, which is divided into five major categories: DOS-Denial of Service (e.g. a mail bomb), R2L- Unauthorized access from a remote machine (e.g. sendmail), U2RUnauthorized access to super user or root functions (e.g. a buffer overflow attack), Probing-surveillance and other probing for vulnerabilities (e.g. port scanning) [13].For each record, KDDCup99 training data set contains 41 fixed feature attributes and a class identifier. In the 41 fixed feature attributes, nine characteristic properties is the discrete type, and others are continuous. In this paper we will use the subset of the original dataset which consist the distinct records. In order to make the data suitable for intrusion detection, we need to preprocess the data. First in order to reduce the number of attribute we apply the information gain algorithm. Second in order to format the dataset we use the Normalization process to normalize the dataset.

## 5.2 Feature Selection

Data Preprocessing is the important task for reducing the attribute of KDD cup 1999 dataset. This process is carried out in two steps. The first step involves mapping symbolic-valued attributes to numeric valued attributes. In second step attributes are reduced by using Information gain. In this first we calculate the entropy of each attribute and subtract the entropy of each attribute by entropy of class label attribute. This calculates the information gain of each attribute. Then we select only those attribute which have positive information gain and other attributes are discarded. The KDD dataset has 41 attributes and after applying information gain 18 attributes remain.

## 5.3 Normalization

A problem with typical data is that different features are on different scales. This cause bias toward some features over other features. To solve this problem, we convert the data instances to a standard form based on the training dataset's Distribution. That is, we make the assumption that the training dataset accurately reflects the range and deviation of feature values of the entire distribution. Normalization also converts the data in the range of 0 and 1 [14]. In Normalization first we select the maximum and minimum value in a particular column and then apply the Normalization Formula given below.

Matrix normalized (j, i) = (selected_column (j)-minimum) / (maximum - minimum)
Where 'j' is the row of matrix
'I' is the column of matrix

## 5.4 Hybrid PSO-SVM for Feature Selection and Parameters

### *5.4.1 Standard Particle Swarm Optimization (SPSO)*

Particle Swarm Optimization was first introduced by Dr. Russell C. Eberhart and Dr. James Kennedy in 1995. Particle Swarm has two primary operators: Velocity update and Position update. During each generation each particle is accelerated toward the particles previous best position and the global best position. At each iteration a new velocity value for each particle is calculated based on its current velocity, the distance from its previous best position, and the distance from the global best position. The new velocity value is then used to calculate the next position of the particle in the search space. This process is then iterated a set number of times or until a minimum error is achieved [15].

Each Particle keep track of its coordinates in the space, which are associated with the best solution the particle has achieved so far. This fitness value is called pbest.

When particle takes the whole population as its topological neighbor, the best value is global "best" value and is called gbest.
SPSO is used to select three parameters C, ε and σ.
C is a cost function
ε - Radial Basis Function
σ – Estimated Accuracy
The PSO algorithm proceeds as follows:
Initialize Population
While (no. of generation)
for p=1 to No. of Particles
If the fitness of Xp is greater than fitness of pbestpthen update pbestp= Xp
for k ε Neighbor of Xp
if the fitness of Xk is greater than that of gbest then update gbest = Xk
Next k
for each dimension d
$V_{pd}^{new} = W * V_{pd}^{old} + C1 * rand1 * (pbestpd - X_{pd}^{old}) + C2 * rand2 * gbestpd - X_{pd}^{old})$
$X_{pd}^{new} = X_{pd}^{Old} + V_{pd}^{new}$
Next d
Next p

Where rand1 and rand2 ε [0, 1]
C1 and C2 = 2

### *5.4.2 Binary PSO*

It is used to feature selection. Dataset with unimportant and noisy feature will decrease the classification Accuracy rate.
For BPSO Algorithm, Xi, Pi and gi for each dimension are between [0, 1]. But this limitation is not for velocity.
For velocity sigmoid function is used.
S (Vpd) = 1/1+exp (-Vpd)
The position is updated by
if rand () < $S(V_{pd}^{old})$ then $X_{pd}^{old} = 1$
else
$X_{pd}^{old} = 0$

The selected features, parameter values and training dataset are used to building SVM Classifier [16].

### *5.4.3 Hybrid PSO-SVM Approach*

Firstly, SPSO is used to elect the C,ε and σin SVM.
Secondly, we selected the best feature subsets by using
BPSO algorithm. The basic process of the PSO algorithm is given by:
Step 1: (Initialization) arbitrarily generate initial particles. For the BPSO algorithm, the complete set of features is represented by a binary string of length N , where a bit in the string is set to '1' if it is to be kept, and set to '0' if it is to be discarded, and N is the original number of features.
Step 2: (Fitness) Measure the fitness of each particle in the population. The selection of this fitness function is a crucial point in using the PSO algorithm, which determines what a PSO should optimize. Here, the task of
The PSO algorithm is to find the global minimum value according to the definition of the fitness function. The definition of the fitness function for the basic method is simply the accuracy of detection.
Step 3: (Update) Compute the velocity of each particle.
Step 4: (Construction) for each particle, move to the next position.

Step 5: (Termination) Stop the algorithm if the termination criterion is satisfied; return to Step 2 otherwise.

# 6. SVM CLASSIFIER

SVM classifier [17] is used to produce better result for binary classification when compared to other classifier. In our proposed technique nonlinear kernel function are used and resulting maximum margin hyper-plane fits in a transformed feature space is a Hilbert space of infinite dimensions. The Gaussian Radial Basics function is given by the equation below.

$$K(X, X') = \exp(- \|X-X'\|2/2\ \sigma 2)$$

The x' defines the center of radial basis function, the vector 'x' is the pattern applied to the input. σIs a measure of width of "x' " Gaussian function with center x'.

The input dataset having large number of attributes is changed into data having k+1 attributes by performing the above steps. The data is given to the SVM to detect if there is any intrusion or not.

# 7. EXPERIMENTAL SETUP AND RESULT

For evaluating the performance of our proposed technique, we had conducted various experiments on KDD Cup 99 dataset. We performed our experiments on Mat lab R2012a on a windows PC with i3 1.80 GHz and 4GB RAM.

## *7.1Data Preparation*

KDD Cup 99 dataset is prepared by MIT Lincoln laboratory. This dataset is publicly available that include actual attacks. For this reason, researchers using this dataset for experiments.

KDD Cup 1999 dataset include both Normal and Malicious attacks and this dataset is obtained from raw TCP dump data for nine weeks. There are about five millions connection records flagged/marked as training data. Each record contain the 41 features/attributes to describing the same connection and is also marked as either normal or a malicious attack. Of the 41 features F(1-9) stands for basic features of a packet, F(10-22) for content , F(23-31) for traffic and F(32-41) for host based features. There are 38 different known attacks in training and test data together, which has been categorized under four category namely Denial of Services (DOS), Probe, remote to local (r2l) and user to root (u2r).

It was found that DOS and PROBE category attack come with greater frequency than other two attacks and can be easily separated from normal activities. It was also found that it became difficult to achieve detection accuracy in dealing with user to root (u2r) and remote to local (r2l) which are embedded in the data portion of the packet.

# 8. EVALUATION PARAMETERS

Evaluation can be done by four parameter True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

True Positive means the attacks which are correctly classified as an attacks.

True Negative means IDS does not make any mistake in spotting a normal condition.

False Positive means attacks which are wrongly classified as attack but they are a valid actions.

False Negative means attacks which are appropriately classified as valid action but they are attacks. A false negative specify that the IDS is incapable to identify the intrusion after a specific attacks has occurred.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Detection Rate} = \frac{TP}{TP+FP}$$

$$\text{False Alarm} = \frac{FP}{FP+TN}$$

Where

FN is False Negative
TN is True Negative
TP is True Positive
FP is False Positive

The detection rate is the number of attacks detected by the system divided by the number of attacks in the data set. The false positive rate is the number of normal connections that are misclassified as attacks divided by the number of normal connections in the data set.

**Table 2: Data points for KDD Dataset**

| **Attack Types** | **Training example** |
|---|---|
| Normal | 12500 |
| Denial of Service | 12500 |
| Remote to user | 39 |
| User to Root | 21 |
| Probing | 1054 |
| Total Attack | 26114 |

The Result are calculated on the Basis of TP, FP, TN and FN. We evaluate the Metrics Namely, Sensitivity, Specificity and Accuracy. From the table it is observed that DOS attack has 99.4% Accuracy and PROBE attack has 99.3% Accuracy. In case of R2L and U2R 98.7% and 98.5% Accuracy respectively.

**Table 3: Accuracy comparison between SVM and Our Technique**

| **Different Method** | **PROBE** | **DOS** | **R2L** | **U2R** |
|---|---|---|---|---|
| SVM | 96.7 | 74.8 | 75.1 | 96.1 |
| Our Proposed Method | 99.3 | 99.4 | 98.7 | 98.5 |

Table 3 Shows that Comparison of Our Technique and Other Technique (SVM). From Table 3 it is clear that our technique is reliable because in the case of DOS attack we have attained 99.4% Accuracy which is Maximum Accuracy. In the case of PROBE we have attained 99.3% Accuracy. For both U2R and R2L we attained a very good accuracy compare to SVM Algorithm which is 98.5 and 98.7%.

**Table 4: Experimental Result**

| **METRICS** | **TYPES OF ATTACKS** | | | |
|---|---|---|---|---|
| | **DOS** | **PROBE** | **U2R** | **R2L** |
| **TRUE NEGATIVE(TN)** | 6642 | 12550 | 12895 | 12572 |
| **FALSE POSITIVE(FP)** | 3 | 83 | 1 | 2 |
| **TRUE POSITIVE(TP)** | 6246 | 444 | 3 | 17 |

| | | | | |
|---|---|---|---|---|
| **FALSE NEGATIVE(FN)** | 192 | 6 | 176 | 492 |
| **Specificity** | 99.9 | 84.2 | 25.0 | 89.4 |
| **Sensitivity** | 97.1 | 99.9 | 98.6 | 96.2 |
| **Accuracy** | 99.4 | 99.3 | 98.5 | 98.7 |

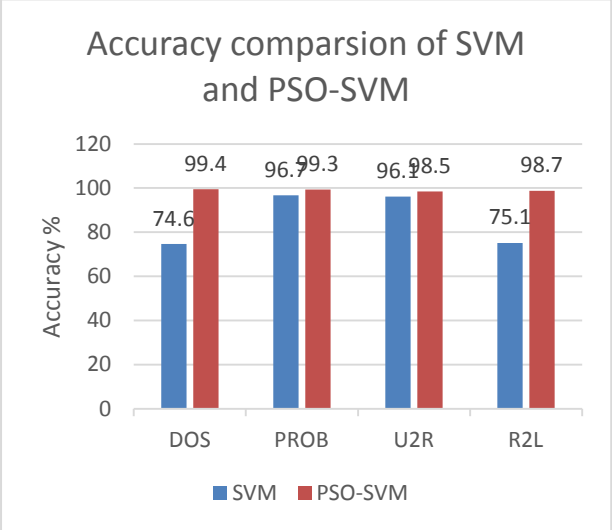


**Figure 2. Graphical Representation of Comparison of Other Technique and proposed technique.**

## 9. CONCLUSION AND FUTURE WORK

Intrusion Detection is a process of detection Intrusion in a computer system in order to increase the security. Intrusion detection is an area in which more and more sensitive data are stored and processed in networked system. We Proposed a Hybrid PSO-SVM approach for building IDS. In SVM parameters C, ε and σ are selected by SPSO. Here we are using two feature reduction technique: Information Gain and BPSO. We analyze that there are several technique which provide good detection rate in case of Denial of Service (DoS) attack. But fail to achieve good detection rate in case of U2R and R2L attack. Many of the algorithm does not perform well in detecting the attacks like U2R and R2L. We perform series of experiment on KDD Cup 99 for acquiring more accuracy. We have used Confusion matrices for evaluation of our proposed technique and the result are obtained on the basis of evaluation metrics namely, Sensitivity, Specificity and Accuracy. As we saw we got the best result as compared to the previous algorithm and it is clear our technique perform well.

## 10. REFERENCES

[1] FengGuorui, ZouXinguo , Wu Jian, "Intrusion Detection Based on the Semi Supervised Fuzzy C-Means clustering Algorithm", Department of information Science and Technology, ShandongUniversity,china , pp. 2667-2670,2012.

[2] Mr. Suresh kashyap ,Ms. Pooja Agrawal, Mr.Vikas Chandra Pandey, Mr. Suraj Prasad Keshri," Soft Computing Based Classification Technique Using KDD 99 Data Set for Intrusion Detection System" in International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering,Vol.2,Issue4,April2013.

[3] R.Durst, T.champion, B.witten, E.Miller, and L.Spagnuolo, "Testing and valuating computer intrusion detection system" communications of ACM, Vo1.42, no.7, pp 53-61, 1999.

[4] Erbacher R F, Walker K L, Frincke D A. Intrusion and Misuse Detection in Large-scale Systems. IEEE Computer Graphics andApplications, 2002, 2(1), pp.38-47.

[5] A.Sung & S.Mukkamala, "Identifying important features for intrusion detection using SVM and neural networks," in symposium on application and the Internet, pp 209-216, 2003.

[6] A.M Chandrasekhar, K.Raghuveer,"Intrusion detection technique by using K-means, Fuzzy Neural Network and SVM classifiers", proceedings of ICCCI, pp1-7, 2013.

[7] Jirapummin, C., Wattanapongsakorn, N., & Kanthamanon, P."Hybrid neural networks for intrusion detection system". Proceedings of ITCCSCC, pp 928-931, 2002.

[8]Horeis, T "Intrusion detection with neural network – Combination of self-organization maps and radial basis function networks for human expert integration", a Research report 2003.

[9] Han, S J & Cho, S. B. "Evolutionary neural networks for anomaly detection based on the behavior of a program", IEEE Transaction on System, Man and Cybernetics, pp 559-570, 2005.

[10] Chen, Y. H., Abraham, A., & Yang, B, "Hybrid flexible neural tree- based intrusion detection systems", International Journal of Intelligent Systems, pp. 337-352,2007.

[11] S. Axelsson, "The base rate fallacy and its implications for the difficulty of Intrusion detection", Proc. Of 6th.ACM conference on computer and communication security 1999.

[12] R.Puttini, Z.marrakchi, and L. Me, "Bayesian classification model for Real time intrusion detection", Proc. of 22nd. International workshop on Bayesian inference and maximum entropy methods in science and engineering, 2002.

[13] A.M Chandrasekhar, K.Raghuveer,"Performance evaluation of data clustering techniques using KDD cup 99 intrusion data set", International journal of information and network security, Vol1(4),pp. 294-305,2012.

[14] Sanjay Kumar Sharma, Pankaj Pandey , Sahel Kumar Tiwari and Mahendra Singh Sisodiya," An Improved Intrusion Detection Based on K-means Clustering via Naïve Bayes Classification", proceedings of ICAESM, pp. 417-422,2012.

[15] Matthew Settles," An Introduction to Particles Swarm Optimization", department of Computer Science, Idaho University.

[16] Jun Wang, XuHong, Rong-rong,Tai-hang Li,"A Real time Intrusion detection system based on PSO-SVM ", proceedings in IWISA Qingdao,China,November 21-22,2009.

[17] MacQueen, Some methods for classification and analysis of Multivariate observations in Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, 1967, pp. 281-297.