A RESULT EVOLUTION OF INTRUSION DETECTION SYSTEM TO IMPROVE THE DETECTION RATE USING ARTIFICAL INTELLIGECE BASED K-MEAN ALGORITHM

Susheel Kumar Tiwari
PhD Research Scholar,
Mewar University, Chittorgarh,
Rajasthan, India

Dr. Manish Shrivastava
Professor & Head (CSE),
L.N.C.T Bhopal, Affiliated to
R.G.P.V Bhopal , India

Abstract: Intrusion Detection Systems are used to monitor computer system for sign of security violations over network or cloud environment. On detection of such sign triggers of IDSs is to report them to generate the alerts. These alerts are presented to a human analyst who evaluates them and initiates an adequate response. In Practice, IDSs have been observed to trigger thousands of alerts per day, most of which are mistakenly triggered by begin events such as false positive. This makes it extremely difficult for the analyst to correctly identify alerts related to attack such as a true positive. Recently Data Mining methods have gained importance in addressing network or cloud security issues, including network intrusion detection and cloud Intrusion detection systems, these systems aim to identify attacks with a high detection rate and a low false alarm rate. Consequently, Unsupervised Learning methods have been given a closer look for network and cloud intrusion detection. We present unsupervised based Clustering Technique and compare with traditional centroid-based clustering algorithms for intrusion detection. These techniques are applied to the KDD Cup98 data set .In addition; a Comparative analysis shows the advantage of proposed approach over Traditional clustering-based Methods over in identifying new or unseen attack. Experimental result show that A.I based Hill Climbing aided k-means Clustering algorithm improves the detection rate in IDS than K-Mean algorithm

KEYWORDS: Intrusion detection system, neural network, false alarm

1. INTRODUCTION

An intrusion detection system (IDS) is a component of the information security framework. Its main goal is to differentiate between normal activities of the system and behavior that can be classified as suspicious or intrusive [1]. The goal of intrusion detection is to build a system which would automatically scan network activity and detect such intrusion attacks. Once an attack is detected, the system administrator can be informed who can take appropriate action to deal with the intrusion**.**

IDS can be host-based (HIDS), network based (NIDS) or a combination of both types (Hybrid Intrusion Detection System). HIDS usually observes logs or system –calls on a single host, while a NIDS typically monitors traffic flows and Network packets on a network segment, and thus observes multiple hosts simultaneously. Generally, one deal with very large volumes of network data, and thus it is difficult and tiresome to classify them manually in order to detect a possible intrusion. One can obtain labelled data by Simulating intrusions, but this will be limited only to the set of known attacks. Therefore, new types of attacks that may occur in future cannot be handled, if those were not part of the training data. Even with manual classification, we are still limited to identifying only the known (at classification time) types of attacks, thus restricting our detection system to identifying only those types. To solve these difficulties, we need a technique for detecting intrusions when our training data is unlabeled, as well as for detecting new and un-known types of intrusions. A method that offers promise in this task is anomaly detection. Anomaly detection detects anomalies in the data (i.e. data instances in the data that deviate from normal or regular ones). It also allows us to detect new types of intrusions, because these new types will, by assumption, be deviations from the normal network usage. It is very difficult, if not impossible, to detect malicious intent of someone who is authorized to use the network and who uses it in a seemingly legitimate way. For example, there is probably no highly reliable way to know whether someone who correctly logged into a system is the intended user of that system, or if the password was stolen.

Under these assumptions we built a system which created clusters from its input data, then automatically labelled clusters as containing either normal or anomalous data instances, and finally used these clusters to classify network data instances as either normal or anomalous. Both the training and testing was done using 10% KDDCup'99 data [2], which is a very popular and widely

used intrusion attack dataset. Most clustering techniques assume a well defined distinction between the clusters so that each pattern can only belong to one cluster at a time[18,19]. This supposition can neglect the natural ability of objects existing in multiple clusters. For this reason and with the aid of fuzzy logic, fuzzy clustering can be employed to overcome the weakness. The membership of a pattern in a given cluster can vary between 0 and 1. In this model a data object belongs to the cluster where it has the highest membership value. In this paper we aim to propose a Neural Network based algorithm which is capable finding unseen attack and identify new attack.

2. RELATED WORK

Security in cloud is one of the major areas of research. The survey shows that, the researchers are focusing on efficient algorithms and encryption techniques to enhance the data security in cloud.

Brian Hay et. al [6] have focused on data authentication, data integrity, querying and outsourcing the encrypted data. Their research says that, the risks can arise at operational trust modes, resource sharing, new attack strategies and digital forensics. In operational trust modes, the encrypted communication channels are used for cloud storage and do the computation on encrypted data which is called as homomorphic encryption. New attack strategies like Virtual Machine Introspection (VMI) can be used at virtualization layer to process and alter the data. The issues are clarified using the digital forensics techniques namely the ephemeral nature of cloud resources and seizing a "system" for examination.

John C. Mace et.al [7] have proposed an automated dynamic and policy-driven approach to choose where to run workflow instances and store data while providing audit data to verify policy compliance and avoid prosecution. They also suggest an automated tool to quantify information security policy implications to help policy-makers form more justifiable and financially beneficial security policy decisions. Service oriented architecture (SOA) is used for work flow deployment in an enterprise. For efficiency, productivity and to achieve public cloud, the cloud computing uses the approaches like retaining control, setting policy, monitoring and runtime security. The dynamic deployment approaches in public cloud computing are security assessment, work flow deployment, policy assignment, audit data and policy analysis.

Qiang Guo et.al [8] gives the unique definition for trust in cloud computing and various issues related to trust are discussed here. An extensible trust evaluation model named ETEC has been proposed which includes a time-variant comprehensive evaluation method for expressing direct trust and a space variant evaluation property for calculating recommendation trust. An algorithm based on ETEC model is also shown here. This model also calculates the trust degree very effectively and reasonably in cloud computing environments.

Bakshi et. al. [9] proposed another cloud intrusion detection solution. The main concern was to protect the cloud from DDoS attacks. The model uses an installed intrusion detection system on the virtual switch and when a DDoS attack is detected.

Xie [10] used Support Vector Machine (SVM) in spam detection. They found two optimal parameters, cost and gamma. They used a good method for selecting proper values of them, which is called "grid search", i.e. to search for the values of certain parameters over supplied parameter ranges. Although they performed parameters optimization, their detection rates were too low. They also did not perform feature selection

3. FUNDAMENTAL THEORY

3.1 Intrusion Detection System

An Intrusion Detection System (IDS) constantly monitors actions in a certain environment and decides whether they are part of a possible hostile attack or a legitimate use of the environment. The environment may be a computer, several computers connected in a network or the network itself. The IDS analyzes various kinds of information about actions emanating from the environment and evaluates the probability that they are symptoms of intrusions. Such information includes, for example, configuration information about the current state of the system, audit information describing the events that occur in the system (e.g., event log in Windows XP), or network traffic. Several measures for evaluating The more widely used measures are the True Positive (TP) rate, that is, the percentage of intrusive actions (e.g., error related pages) detected by the system, False Positive (FP) rate which is the percentage of normal actions (e.g., pages viewed by normal users) the system incorrectly identifies as intrusive, and Accuracy which is the percentage of alarms found to represent abnormal behavior out of the total number of alarms. In the current research TP, FP and Accuracy measures were adopted to evaluate the performance of the new methodology.

3.2 Network Profiling

Since the number of attacks is always increasing, IDS should be updated with signature for new attacks. Network profiling can help IDS to define labels of new signatures. There are some problems in network profiling such as grouping the attacks that come through the network based on their types. Those problems can be solved using data mining techniques such as clustering and classification [14,15]

3.3 Clustering Techniques

Cluster analysis is the process of partitioning data objects (records, documents, etc.) into meaningful groups or clusters so that objects within a cluster have similar characteristics but are dissimilar to objects in other clusters. Clustering can be viewed as unsupervised classification of unlabelled patterns (observations, data items or feature vectors), since no pre-defined category labels are associated with the objects in the training set. Clustering results in a compact representation of large data sets (e.g., collections of visited Web pages) by a small number of cluster centroids. Applications of clustering include data mining, document retrieval, image segmentation, and pattern classification

3.4 Classification

Classification is the task of assigning objects to one of several categories. A classification model can predict the class label of unknown object. Classification often used in biology and financial. In classification, datasets are divided into search domain and new sample. Classification technique builds a classification model from the search domain and decide the class label for each given input/object. Some classification algorithms are -Nearest Neighbor, Decision Tree, and Support Vector Machine (SVM)[10].

4. PROPOSED APPROACH

We propose Artificial Intelligence based clustering algorithm for network intrusion detection. This k-means algorithm aims at minimizing a squared error function is given in Equation for the objective function.

$$J = \sum_{i=1}^{k} \sum_{i=1}^{n} \left\| x_i(j) - c_j \right\|^2$$

Where $\left\| x_{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point xj (j) and the cluster centre cj is an indicator of the distance of the n data points from their respective cluster centers. One of the main disadvantages to K-Mean algorithm is that it requires the number of clusters as an input to the algorithm. The algorithm is incapable of determining the appropriate number of clusters and depends upon the user to identify this in beforehand. For example, if you had a group of people that were easily clustered based upon gender while calling the k-means algorithm with k=3 would force the people into three clusters and when k=2 would provide a more natural fit. Likewise, if a group of individuals were easily clustered based upon home state and you called the k-means algorithm with k=20 then the results might be too generalized to be effective.

But finding the value of i that best suits of data is very difficult. Hence we moved on to hill climbing. Hill climbing is good for finding a local optimum (a good solution that lies relatively near the initial solution) but it is not guaranteed to find the best possible solution (global optimum) out of all possible solutions (search space) which can be overcome by using steepest ascent Modified Hill climbing finds globally optimal solution. The relative simplicity of the algorithm makes it a popular first choice amongst optimizing algorithms and it is widely used in artificial intelligence, in order to reach a good state from a start state. Selection of next node and starting node can be varied to give a list of related algorithms. This can often produce a better result than other algorithms when the amount of time available to perform a search is limited, such as with real-time systems. Artificial Intelligence approach based Hill climbing algorithm attempts to maximize (or minimize) a target function $f(x)$ where x is a vector of continuous and / or discrete values. In each iteration, hill climbing will adjust a single element in x and determine whether the change improves the value of $f(x)$. Then, x is said to be globally optimal

Artificial Intelligence approach based Hill Climbing aided k-means Algorithm steps are shown bellow.

Input: randk - random value of kΔk - A random move in cluster

Output: k - Number of clusters Pseudo code: Modified Hill Climbing Algorithm

do

l1: iter =true;

ksolved ← randk;

l2: newsolution ← ksolved + Δk;

if (f (newsolution) < f (ksolved) then

solution ← newsolution;

ksolved ← solution; k←ksolved;

if (algorithm converged and globally optimum) then

output k;

iter = false;

else goto l2 ;

else goto l1 ;

while (iter);

Input: E= { e1, e2…en } - Set of entities to be clustered

k - number of cluster from Modified Hill Climbing Algorithm MaxIters - Limit of iterations

Output: C= {c1, c2…cn } - Set of clustered

centroids

L= {l (e) e= {1, 2…n} - Set of cluster labels of E

Pseudo code:

Modified Hill Climbing aided k-means Algorithm

for each ci ϵ C

do ci ← ej ϵ E (E.g. random selection);

end

for each ei ϵ E do

L (ei) ← argmin Distance (ei, ci)j ϵ {1,…, k};

end changed ← false;

iter ← 0; repeat

for each ci ϵ C do

Update cluster (ci);

End

for each ei ϵ E do

minDist ← argminDistance (ei ,cj) jϵ {1…k};

if minDist ≠ l (ei) then;

l(ei) ← minDist;

changed ← true;

end

end

iter ← iter+1;

until changed=true and iter ≤ MaxIters;

In the above algorithm is the best K value is obtained by modified hill climbing and this value is utilized in k– means algorithm in order to form effective clusters with uniform cluster density. The following section deals with performance evaluation of implemented system

5. EVALUATION MEASURES

To evaluate the system performance the following measures (based on Sequeira and Zaki 2002) were used.

True Positive Rate (TP) (also known as Detection Rate or Completeness): the percentage of terrorist pages receiving a rating above the threshold in the experiments, terrorist pages will be obtained from the users simulating terrorists.

False Positive Rate (FP): the percentage of regular Internet access pages that the system incorrectly determined as related to terrorist activities, i.e., the percentage of non-terrorist pages receiving a rating above threshold and suspected falsely as terrorists.

Accuracy: percentage of alarms related to terrorist behavior out of the total number of alarms. Since no benchmark data on content based intrusion detection is currently available, the results are compared to the best numbers achieved with ADMIT which is a command level method using the Means clustering algorithm to detect intruders.

6. EXPERIMENTAL RESULT

6.1 Data Set Description

Because the goal of this work is to study and enhance the learning capabilities of the Artificial Immune System techniques for intrusions detection, A.I based Hill Climbing aided k-means Clustering algorithm is compared to a clustering based k-mean algorithm that use the full set of samples sampled from the KDD Cup98 dataset and witch contain 5000 sample. The original data set contain 5 million records which specify various attacks in which 1% sample consisting of about 5000 records was used in our experiment.

6.2 Data Preprocessing

In K-DD-98 data set, each records representing a connection between two networks host according to some well defined network protocol. Each connection is represented by 41 features, which include the basic features of individual of TCP Connections, the content features, No. of byte, Transferred byte etc. the features in column 2 in KDD-98 Data set are transmitted byte, flag. We are interested in anomaly detection via unsupervised Learning algorithm. Hence all the records labeled as attack were considered as intrusion, while remaining was considered as normal. The labels are not used during clustering process but are used for evaluating the detection performance of the algorithm. The clustering algorithm used in our comparative study do not handle categorical data the categorical feature such as flag in the data set are converted using 1-to-N Encoding Scheme.

6.3 Empirical Setting

The K-Means[16,17] and Artificial Intelligence based Clustering algorithm are written in visual basic.Net 2008 as front-end and MS-Access used as Backend and compiled into mix files. Artificial Intelligence based Clustering algorithm are relatively efficient due to vectored programming and active optimization.

In order to study the effect of the total number of clusters on the intrusion detection results, we performed empirical studies with 100 total numbers of clusters.

For clustering quality, we use the Computation time. The purity of a cluster is defined as the percentage of the most dominated instance category in the cluster, and average purity is the mean over all clusters. Its value can range from 0 to 1, with higher values representing better average purity. The run time of each algorithm is also recorded and compared. Each experiment is run five times for evaluating intrusion detection results, we report the detection rate .The false positive rate is the percentage of normal instances that are labeled as attacks. The attack detection rate represents the percentage of all attack instances that are detected, i.e., labeled as attacks. We also report in Graph, by varying the parameter used in the detection method, to show the tradeoff between the false positive rate and the detection rate.

7. EXPERIMENT ANALYSIS

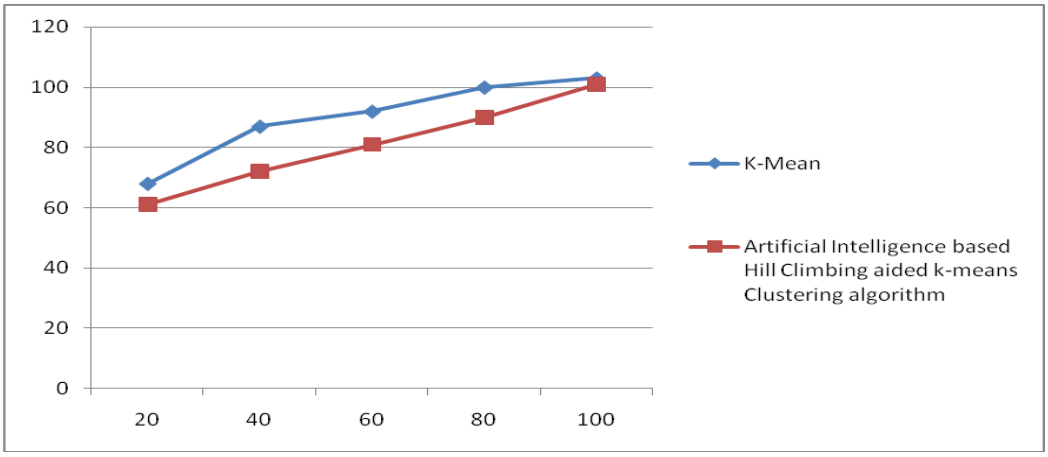
Now, we compare the aforementioned clustering algorithms on the whole data set with 5000 data set The Computation time for the clustering algorithms with 100 clusters are shown in Table 1 respectively.

Experiment 1:

In this experiment 1, we investigate computation time of Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm and K-Mean algorithm. In the training phase, the Selection Algorithm was used to cluster the training data. After training, each cluster was labeled according to the majority type of data in this cluster. For instance, if more than 50% of the connections in cluster were intrusions, the cluster and its centroid weight vector would be labeled as intrusion. Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm perform significantly better ($p < 5\%$) than the others in terms of computation time with much less run time.

Comparing the results for 100 clusters is shown in table (1). Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm algorithms perform significantly better ($p < 5\%$) than the others in terms of computation time. Comparing the results for 100 clusters, we observe that the K-Means take more execution time than Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm. This experiment is run on individual clusters for each individual cluster on KDD98 Data set. This data set contain only numeric value not categorical valued. Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm is fast than K-Mean

Table(1): Computation Time of K-Mean and Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm



Fig(.1): Clustering results with 100 clusters with time efficiency

Cluster	Algorithm	
	K-Mean	Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm
	Time (ms)	Time (ms)
20	68	61
40	87	72
60	92	81
80	100	90
100	103	101

Since our aim is to detect network intrusion using clustering algorithms [10], we now analyze the unsupervised intrusion detection accuracies or times for detect the unseen or new attack. We sort clusters according to their possibility of being normal in decreasing order and arrange data instances in a cluster in the same way. The distance to the centroid of the largest cluster measures the possibility of being normal.

Graph can be constructed by dividing the sorted data instances into normal and intrusive categories at a series of cutting points. Figure(.1) shows that the decline of time is fast when the value of k is very small, since the instances in the same cluster may be quite different from each other Splitting clusters can significantly decrease the value of time after k reaches the turning point, the decline of time will become slow. At this point, the data set may have been well partitioned.

Experiment 2:

Now we find the detection rate of K-Mean and Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm. To evaluate the accuracy of a system, we use two indicators, which were used in: Detection Rate (DR) and False Alarm Rate (FAR).
DR equals the number of intrusions divided by the total number of intrusions in the data set

$$DR = \frac{\text{the number of intrusions}}{\text{The total number of intrusions in the data set}}$$

We partitioned 5000 instances of KDD-99 data using the K-Mean Algorithm and Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm with different initial values of k. The Detection rate of K-Mean algorithm and A.I based Hill Climbing aided k-means Clustering algorithm are

Cluster	K-Mean algorithm	Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm
20	85	90
40	75	80
60	69	76
80	63	73
100	61	72

Table (2): Summary Detection results with 100 clusters

The table (2) shows that K-Mean algorithm has low detection rate than A.I based Hill Climbing aided k-means Clustering algorithm. Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm is used in intrusion detection system is so good when it has low false positive rate and high Detection rate. Proposed Algorithm map reduce the false positive and it has high detection rate for detect the unseen or new attack. The graph for the K-Means and Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm are omitted for comprehensibility and better visualization, particularly because they are visibly worse.

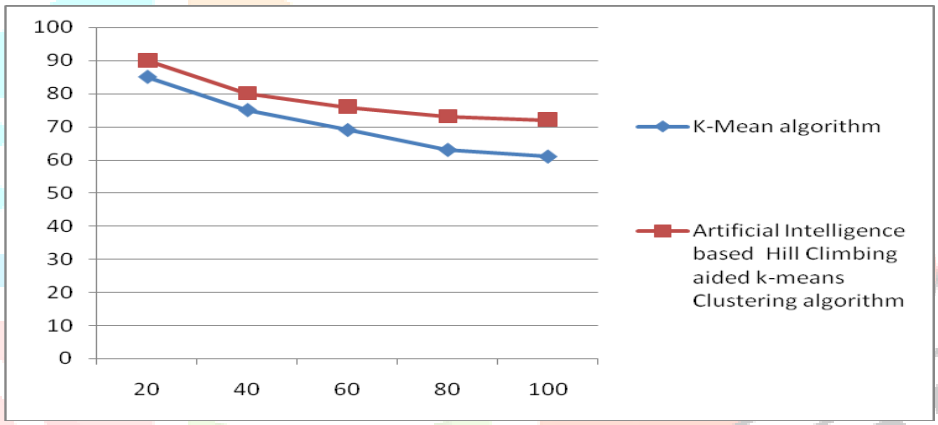


Fig. (2) : Detection rate of the K-Mean algorithms and Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm

Fig. (2) Show the Graph for detection rate of the K-Mean algorithms and Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm with 100 clusters respectively. It can be seen that for 100 clusters, the A.I based Hill Climbing aided k-means Clustering algorithm has high detection rate than K-Mean.

8. CONCLUSION

An approach for a neural network based intrusion detection system, intended to classify the normal and attack patterns and the type of the attack, has been presented in this paper. We applied Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm method which increased the generalization capability of the neural network and at the same time decreased the training time. It should be mentioned that the long training time of the neural network was mostly due to the huge number of training vectors of computation facilities. However, when the neural network parameters were determined by training, classification of a single record was done in a negligible time. Therefore, the neural network based IDS can operate as an online classifier for the attack types that it has been trained for. The only factor that makes the neural network off-line is the time used for gathering information necessary to compute the features. In this paper we introduce Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm in a cloud to protect user. This algorithm is intended to be scalable by allowing different format of data to apply into Artificial Intelligence based Hill Climbing aided k-means Clustering algorithm for more reliable IDS solution.

REFRENCES

1. Mahesh s,Mahesh T R, M Vinayababu, " Using Data Mining Techniques for Detecting terror related activities on the web", Journal of Theoretical and Applied information technology,2016

2. Abbasi, A., & Chen, H. Applying authorship analysis to extremist group Web forum messages. IEEE Intelligent Systems, Special Issue on Artificial Intelligence for National and Homeland Security, 20(5), 67–75,2016
3. Baumes, J., Goldberg, M., Hayvanovych, M., Magdon-Ismail, M., Wallace, W., & Zaki, M. Finding hidden group structure in a stream of communications. In S. Mehrotra, D.D. Zeng, & H. Chen (Eds.), Proceedings of the IEEE Conference on Intelligence and Security Informatics (pp. 201–212). Los Alamitos, CA: IEEE.,2016
4. Chen, H. Intelligence and security informatics: Information systems perspective. Decision Support Systems: Special Issue on Intelligence and Security Informatics, 41(3), 555–559.,2005
5. Chen, H., Qin, J., Reid, E., Chung, W., Zhou, Y., Xi, W., et al. (2015). The dark Web portal: Collecting and analyzing the presence of domestic and international terrorist groups on the Web. In W.T. Scherer & B.L. Smith (Eds.), Proceedings of the 7th IEEE International Conference on Intelligent Transportation Systems, (pp. 106–111).
6. Brian Hay, Kara Nance, Matt Bishop, "Storm Clouds Rising: Security Challenges for IaaS Cloud Computing" Proceedings of the 44th Hawaii International Conference on System Sciences -2016.
7. John C.Mace, Aad van Moorsel, Paul Watson, "The Case for Dynamic Security Solutions in Public Cloud Workflow Deployments" School of Computing Science & Centre for Cybercrime and Computer Security (CCCS) Newcastle University, Newcastle upon Tyne, NE1 7RU, UK,2016.
8. Qiang Guo, Dawei Sun, Guiran Chang, Lina Sun, Xingwei Wang, "Modeling and Evaluation of Trust in Cloud Computing Environments" School of lnformation Science and Engineering, Northeastern University, Shenyang, P.R. China, Computing Center, Northeastern University, Shenyang, P.R. China, 2011 3rd International Conference on Advanced Computer Control (ICACC 2015).
9. *Bakshi,* Chun-Chieh Huang, Joy Ku, "A Cooperative Intrusion Detection System Framework for Cloud Computing Networks", 39th International Conference on Parallel Processing Workshops, 2016.
10. Y. Xie. An Introduction to Support Vector Machine and Implementation in R. May 2015.
11. Johan Zeb Shah and anomie bt Salim, "Fuzzy clustering algorithms and their application to chemical datasets", in Proc. Of the post graduate Annual Research seminar 2015, pp.36-40.