Intrusion Detection System by using K-Means clustering, C 4.5, FNN, SVM classifier

Miss Meghana Solanki[1], Mrs. Vidya Dhamdhere[2]

PG Student [1], Faculty [2]
G. H. Raisoni , college of Engineering & Management , Wagholi , Pune

Abstract

Security of Information is one of the cornerstones of Information Society. As number of Network attacks have increased over the past few years, intrusion detection system (IDS) is increasingly becoming a critical component to protect the network. .In recent years, many researchers are using data mining techniques for building IDS. One of the main challenges in the security management of large-scale high-speed networks is the detection of suspicious anomalies in network traffic patterns due to Distributed Denial of Service (DDoS) attacks or worm propagation. Intrusion detection methods started appearing in the last few years. Here, We present a Intrusion detection method using K-means clustering, neuro-fuzzy models, Support vector machine (SVM) and C4.5 algorithm. We propose a four level framework for Intrusion detection in which first procedure concerns to generate different training subsets by using k-means clustering, second procedure based on the training subsets different neuro-fuzzy models are trained, third procedure a vector for SVM classification and radial SVM classification is perform. Finally we build the decision tree using C4.5 decision tree algorithm.

Keywords:- Fuzzy Neural Network ,Intrusion Detection System ,Network Intrusion Detection System ,SVM etc

1.INTRODUCTION

In today's scenario network is most essential part of the communication. Individual can do Lot of things on internet. Internet gives many advantages and but has some disadvantages too. It is used as a one of the tool for crime. One of the major and famous crimes is hacking. Network security has become a critical issue owing to the incredible growth of computer networks usage. It becomes technically hard and economically expensive for the manufactures to secure the computer systems from external attacks. A Network Intrusion Detection System (NIDS) is a device (or application) that examines network and/or system activities for malicious activities or policy violations and produces reports to a Management Station. Intrusion detection is the process of monitoring the events occurring in a computer system or network. It is used for analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices. Intrusion detection system detects attacks and anomalies in the network, and thus are becoming very important. IDS are useful in detecting successful intrusion, and also in monitoring the network traffic and the attempts to break the security. Intrusion detection is the practice of observing and examining the actions going on in a system in order to identify the attacks and susceptibilities. Data mining is the modern technique for analysis of huge of data such as KDD CUP 99 data set that is applied in network intrusion detection. Large amount of data can be handled with the data mining technology. It is still in developing state, as it is growing rapidly it can become more effective. Intrusion detection methods started appearing in the last few years. Using intrusion detection methods, you can collect and use information from known types of attacks and find out if someone is strying to attack your network or particular hosts. The information collected this way can be used to harden your network security, as well as for legal purposes. Both commercial and open source products are now available for this purpose. Many vulnerability assessment tools are also available in the market that can be used to assess different types of security holes present in your network.

2. LITERATURE SURVEY

Zhi-song pan et al., [2003] have reported a misuse intrusion detection model based on a hybrid neural network and decision tree algorithm. They have discussed the advantages of different classification abilities of neural networks and the C4.5 algorithm for different attacks [1]. While neural network algorithm is reported to have high performance to DOS and Probe attacks, the C4.5 algorithm has been found detect R2L and U2R attacks more accurately. In general, IDSs can be divided into two techniques: misuse detection and anomaly detection [2], [3]. Misuse detection refers to detection of intrusions that follow well-defined intrusion patterns. It is very useful in detection known attack patterns. Anomaly detection refers to detection performed by detecting changes in the patterns of utilization or behavior of the system. It can be used to detect known and unknown attack. The anomaly detection techniques have the advantage of detecting unknown attacks over the misuse detection technique [4]. Anomaly based intrusion detection using data mining algorithms such as decision tree (DT), naïve Bayesian classifier (NB), neural network (NN), support vector machine (SVM), k-nearest neighbors (KNN), fuzzy logic model, and genetic algorithm have been widely used by researchers to improve the performance of IDS [5][6].This paper present an efficient technique for intrusion detection by making use of k-means clustering, fuzzy neural network and radial support vector machine. System uses different techniques for intrusion detection [12]. In this paper an intrusion detection system is developed using

Bayesian probability. The system developed is a naive Bayesian classifier that is used to identify possible intrusions [13]. In this paper we propose a new, quantitative-based approach for the detection and the prevention of intrusions. Our model is able to probabilistically predict attacks before their completion by using a quantitative Markov model built from a corpus of network traffic collected on a *honey pot* [14]. Paper focus on improving intrusion system in wireless local area network by using Support Vector Machines (SVM). SVM performs intrusion detection based on recognized attack patterns [15].

3. RELATED WORK

A. Hybrid Approach

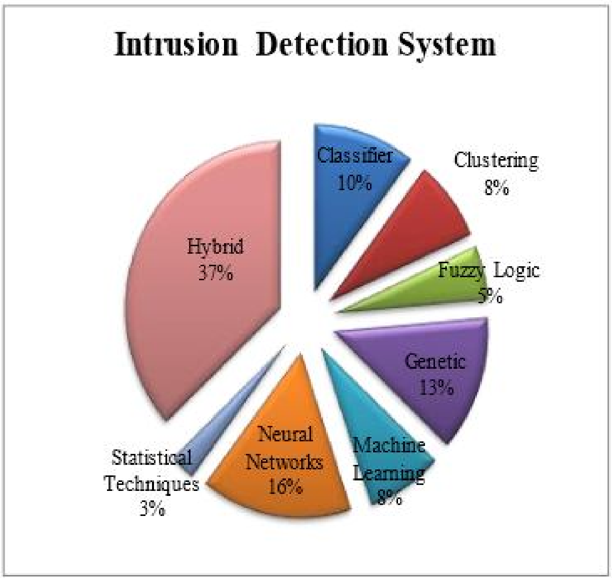

Fig 1 the percentage distribution of the number of method under various IDS approaches

Fig 1 shows Hybrid approach improves the accuracy of the IDS when compared to single approaches. Results from the different individual systems are combined to provide more accuracy and reliability. Researchers are focusing on hybrid methodology for developing the IDS as it can combine the advantages of two algorithms.

B.Classification of Intrusion Detection and Intrusion Detection Systems

I) MISUSE DETECTION SYSTEM

Misuse detection is an approach in detecting attacks. In misuse detection approach, we define abnormal system behavior at first, and then define any other behavior, as normal behavior. It stands against anomaly detection approach which utilizes the reverse approach, defining normal system behavior and defining any other behavior as abnormal. In other words anything we don't know is normal. Using attack signatures in IDSes is an example of this approach. Misuse detection has also been used to refer to all kinds of computer misuse.

II) Network Based Intrusion Detection System

Network based Intrusion Detection System (NIDS) monitors the traffic as it flows to other host. Monitoring criteria for a specific host in the network can be increased or decreased with relative ease. NIDS should be capable of standing against large amount of network traffic to remain effective. As network traffic increases exponentially NIDS must grab all the traffic and analyze in a timely manner.

III) Anomaly Based Intrusion Detection System

Anomaly based Intrusion Detection System examines ongoing traffic, activity, transactions and behavior in order to identify intrusions by detecting anomalies. It works on the notion that attack behavior differs enough from normal user behavior such that it can be detected by cataloging and identifying the differences involved. The system administrator defines the baseline of normal behavior. Anomaly-based IDS systems are very prone to a lot of false positives .Anomaly-based IDS systems can cause heavy processing overheads on the computer system. In this case we have two possibilities: (1)False positive: Anomalous activities that are not intrusive but are flagged as intrusive. (2) False Negative: Anomalous activities that are intrusive but are flagged as non intrusive. The block diagram of anomaly detection system is as following:

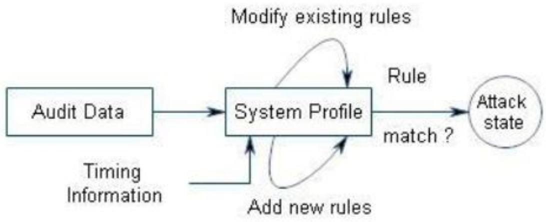

Fig 3 Misuse Detection Systems

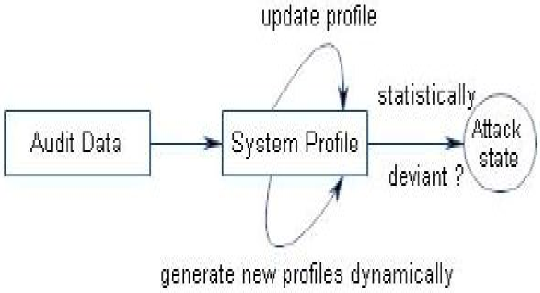

Fig 4 Anomaly Detection Systems

C .Methodology OF IDS

Clustering Techniques: Clustering is the practice of combining the records into classes or clusters, so that entities within the cluster have high resemblance in compare to one another on the further hand are very dissimilar to entities in other clusters. There are alternate approaches which are used in clustering process Partitioning methods, Hierarchical methods, etc.

Fuzzy Logic: Fuzzy Logic is a problem solving control Structure approach that gives itself to implementation in the systems which are ranging from multichannel PC or Workstation acquisition and control systems. It can be engaged in hardware, software, or in both. It offers a simple manner to attain on a definite decision based upon indefinite, ambiguous, inaccurate, noisy, or absent input information

Support Vector Machine (SVM)

Support vector machines (SVM) are learning machines that plot the training vectors in high dimensional feature space, labeling each vector by its class. SVMs classify data by determining a set of support vectors, which are members of the set of training inputs that outline a hyper plane in feature space. Computing the hyper plane to separate the data points leads to a quadratic optimization problem. There are two main reasons that we used SVMs for intrusion detection. The first reason is that its performance is in terms of execution speed, and the second reason is scalability. SVMs are relatively insensitive to the number of data points, and the classification complexity does not depend on the dimensionality of the feature space [11].

K -Means Clustering

The K-means algorithm, starting with k arbitrary cluster centers in space, partitions the set of giving objects into k subsets based on a distance metric. The centers of clusters are iteratively updated based on the optimization of an objective function. This method is one of the most popular clustering techniques, which are used widely, since it is easy to be implemented very efficiently with linear time complexity[12]. The principle goal of employing the K-Means clustering scheme is to separate the collection of normal and attack data that behave similarly into several partitions which is known as *K*-th cluster centroids. In other words, K-Means estimates a fixed number of *K*, the best cluster centroid representing data with similar behavior. In our work, we predefined K=3,representing Cluster 1, Cluster 2, and Cluster 3. Thus, the iterative K-Means algorithm is designed as follows:

Initially: Randomly select K = 3 cluster centroid .Do Correspond data point to nearest clusters. Update optimal cluster centroid based on corresponding data points and labeling the points While no change remains Certain. activities or data are alike to either normal or abnormal behavior. The K-Means algorithm is unable to differentiate this behavior r precisely.

Intrusion detection with k-means clustering and c4.5 decision tree learning methods

k-Means clustering and the C4.5 decision tree classification methods for supervised anomaly detection.

Intrusion Detection with k-Means Clustering

The k-Means algorithm groups N data points into k disjoint clusters, where k is a predefined parameter. The steps in the k-Means clustering-based Intrusion detection method are as follows:

Step 1: Select k random instances from the training data subset as the centroids of the clusters C1; C2;...Ck.

Step 2: For each training instance X:

a. Compute the Euclidean distance *D(Ci,X),i = 1...k*

b. Find cluster Cq that is closest to X.

c. Assign X to Cq. Update the centroid of Cq. (The centroid of a cluster is the arithmetic mean of the instances in the cluster.)

Step 3: Repeat Step 2 until the centroids of clusters C1; C2; ...Ck stabilize in terms of mean-squared error criterion.

Step 4: For each test instance Z:

a. Compute the Euclidean distance *D(Ci,Z),i = 1...k*. Find cluster Cr that is closest to Z.

b. Classify Z as an anomaly or a normal instance using the Decision tree[7].

Intrusion Detection with C4.5 Decision Trees

In C4.5 first grows an initial tree using the divide-and-conquer algorithm as follows:

1. If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S.

2. Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S1, S2,... according to the outcome for each case, and apply the same procedure recursively to each subset[11].

Intrusion Detection with K-MEANS+C4.5

Intrusion detection system can be used by cascading K–Means and C4.5 decision tree algorithm. The proposed method is divided into two phases,

1) Training Phase and 2)Testing Phase.

1. Training Phase i Z

We are provided with a training data set (*,i i X Y*), *i* 1,2,3,......,*N* where *i X* represents an n dimensional continuous valued vector and {0,1} *i Y* represents the corresponding class label with “0”for normal and “1” for anomaly. The method has two steps: 1) training and 2) testing. During training, steps 1-3 of the k-Means-based anomaly detection method are first applied to partition the training space into k disjoint clusters 1 2 3 , , ,......, *K C C C C* . Then, C4.5 decision tree is trained with the instances in each k-Means cluster. The k-Means method ensures that each training instance is associated with only one cluster. However, if there are any subgroups or overlaps within a cluster, the C4.5 decision tree trained on that cluster refines the decision boundaries by partitioning the instances with a set of if then rules over the feature space.

2. Testing Phase

In the testing phase, we have two subdivided phases 1) Selection Phase and 2) Classification Phase. In selection phase, compute the Euclidean distance for every testing instance and find the closest cluster. Compute the decision tree for the closest cluster. In classification phase, apply the test instance *i Z* over the C4.5 decision tree of the computed closest cluster and classify the test instance *i Z* as normal or anomaly. The algorithm for the proposed method is given below[7].

K.Means+C4.5 Algorithm

Selection Phase

Input: Test instances i Z , i 1,2,3,......,N .

Output: Closest cluster to the test instance i Z .

Procedure Selection

Begin

Step 1: For each test instance *i Z*

```
a. Compute the Euclidean distance D(Zi, rj),j=1...k, and find the cluster closest toi Z .
b. Compute the C4.5 Decision tree for the closest cluster.
End /*End Procedure*/
```

Classification Phase

```
Input: Test instance i Z .
Output: Classified test instance i Z as normal or anomaly
Procedure Classification
Begin
Step 1: Apply the test instance i Z over the C4.5 decision tree of the computed
closest cluster.
Step 2: Classify the test instance i Z as normal or anomaly and include it in the cluster.
Step 3: Update the centre of the cluster.
End /*End Procedure*/
```

Architecture of IDS

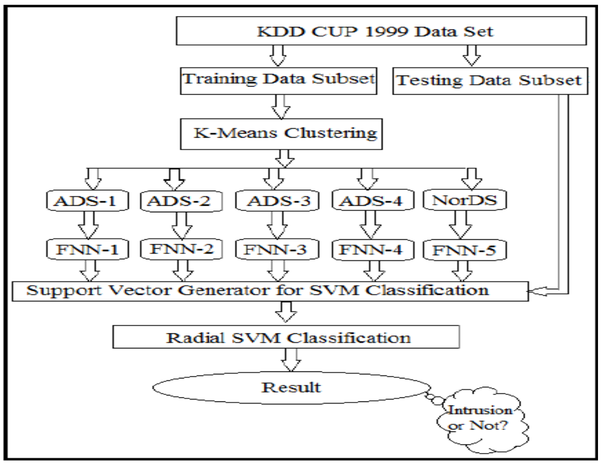

Fig 5. Architecture for IDS

The architecture of technique as shown in the fig 5 consists of 5 step methodology. This can be explained as follows.

1. The input data set DS needed for experimentation is prepared by conducting relevance analysis on KDD Cup 1999 data set in order to reduce the irrelevant attributes / features which will not contribute for intrusion detection.

2. The input dataset is divided into Training Data set and Testing Data set. The Training data is clustered using K-Means Clustering into k subsets where, k is the number of clusters desired.

3. Neuro-fuzzy (FNN) training is given to each of the k cluster, where each of the data in a particular cluster is trained with the respective neural network associated with each of the cluster.

4. Generation of vector for SVM classification, S={D1, D2,…….DN} which consists of attribute values obtained by passing each of the data through all of the trained Neuro-fuzzy classifiers, and an additional attribute µij which has membership value of each of the data.

5. Classification using SVM to detect intrusion has happened or not.

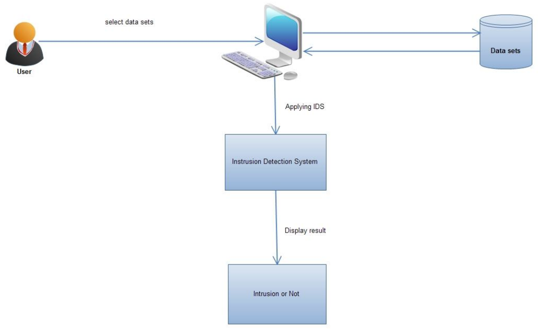

Fig 6 System architecture

fig 6 shows the system architecture. it consist of users, data set & intrusion detection system. . By using data set we determine whether intrusion has occurred or not.

4. Conclusion

Intrusion detection systems (IDSs) play an important role in protecting computer information. IDS users depend on the IDS to protect their computers and networks demand that an IDS provides reliable and continuous detection service. Many of the today's intrusion detection methods generate high false positives and negatives Network Anomaly detection systems are designed based on availability of data instances. Many intrusion detection techniques have been specifically developed for certain application domains, while others are more generic. In this paper, we present a cascaded algorithm using K–Means and C4.5. In this paper performance analysis is measured by using five measures, 1) detection accuracy (or) True Positive Rate(TTR), 2)False Positive Rate (FPR), 3) Precision, 4) Total Accuracy (TA), and 5) F–Measures (FM). The proposed algorithm gives impressive detection accuracy in the experiment results.

References

[1] Zhi-Song Pan, Song-Can Chen, Gen-Bao Hu, Dao-Qiang Zhang, (2010), "Hybrid Neural Network and C4.5 for Misuse Detection " , Proceedings of the second International conference on Machine Learning and Cybernetics, November, pp. 2463 – 2467.

[2] E.Biermann, E. Cloeteand L.M.Venter,"A comparison of intrusion detection Systems", Computer and Security, vol.20, pp.676-683, 2001.

[3] T.Verwoerd and R.Hunt, "Intrusion detection techniques and approaches",Computer Communications, vol.25, pp.1356-1365, 2002.

[4] E. Lundin and E. Jonsson, "Anomaly-based intrusion detection: privacy concerns and other problems", Computer Networks, vol.34, pp.623-640, 2002.

[5] Barbara, Daniel, Couto, Julia, Jajodia, Sushil, Popyack, Leonard, Wu, and Ningning, "ADAM:

Detecting intrusion by data mining," IEEE Workshop on Information Assurance and Security, West Point, New York, June 5-6, 2001.

[6] T. Shon, J. Seo, and J.Moon, "SVM approach with a genetic algorithm for network intrusion detection," In Proc. of 20th International Symposium on Computer and Information Sciences (ISCIS 2005), Berlin: Springer-verlag, 2005, pp. 224-233.

[7] M. Su, "Real time anomaly detection systems for Denial of Service attacks by weighted k nearest neighbor classifiers", Expert Systems with Applications, 2011, 38: p.3492 3498

[8] G. Wang, J. Hao, J. Ma, L. Huang," A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering", Expert Systems with Applications,2010,37: p. 6225 6232.

[9] S. Horng, M. Su, Y. Chen, T. Kao, R. Chen, J. Lai, C.
Perkasa, "A novel intrusion detection system based on hierarchical clustering and support vector machines", Expert Systems with Applications,2011,38: p. 306 313

[10] A. N. Toosi, M. Kahani, "A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers", Computer Communications, 2007, 30: p. 2201–2212.

[11] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, K. Dai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method", Expert Systems with Applications,2011,39: p. 424 430.

[12] M. Panda and M.R. Patra. 2007. Network intrusion detection using naive bayes. IJCSNS International Journal of Computer Science and Network Security, 7, 258-263.

[13] A.M. Chandrasekhar, "Intrusion Detection Technique By Using K-Means, Fuzzy Neural And Svm Classifier ", 2013 International Conference on Computer Communication and Informatics (ICCCI - 2013), Jan 04-06, 2013 Coimbatore, India.

[14] Hesham Altwaijry, "Bayesian Based Intrusion Detection System ", Journal of King Saud University – Computer and Information Sciences (2012) 24,1–6.

[15] Ammar Boulaiche, "A Quantitative Approach For Intrusions Detection And Prevention Based On Statistical N-Gram Models ", Procedia Computer Science 10 (2012) 450 – 457.

[16] Muamer N.Mohammed, "Intrusion Detection System Based On Svm For Wlan ",Procedia Technology 1 (2012) 313 – 317.