ISSN 2395-1621

High Detection Rate by Using Anomaly Based Machine Learning Algorithm

[#1]Prashant Wakhare, [#2]Dr S.T.Singh

[1]Prashant_mitr@rediffmail.com
[2]stsingh47@gmail.com

[#1]Computer Engineering,
Savitribai Phule Pune University
P K Technical Campus, Pune
[#2]Savitribai Phule Pune University
P K Technical Campus, Pune

ABSTRACT

Intrusion detection is a process of identifying the Novel as well as Misuse attacks. The main aim of IDS is to identify the Normal and Abnormal activities. In Previous years, many publishers are using data mining techniques for building Intrusion Detection System. Here we propose a new approach using data mining technique such as k-means, Neuro-fuzzy and radial basis support vector machine for helping IDS to attain higher detection rate. This paper we will implemented few techniques has steps: Preprocessing, clustering using K-means to generate different training subsets then based on training data Subset, different Neuro fuzzy model are trained. Then based on the subsequent training subsets a vector for SVM (Support Vector Machine) classification is formed and at the end, radial SVM is performed to detect intrusion is happen or not. The result of experiment on KDD (Knowledge Discovery Data) CUP 99 dataset is demonstrated. Result shows that BPNN (Back Propagation Neural Network), multiclass SVM and other well-known method- decision trees and Columbia model in terms of sensitivity, specificity and in particular detection accuracy.

Keywords— Intrusion Detection System, K-means clustering, Fuzzy Neural Network, Radial SVM

ARTICLE INFO

Article History

Received : 18th July 2015

Received in revised form :

20th July 2015

Accepted : 24th July 2015

Published online :

27th July 2015

I. INTRODUCTION

The main aim of the Intrusion Detection System (IDS) is to identifying the computer system from attack. The IDS is the most important part of the security infrastructure for the networks connected to the internet because various ways to compromise the stability and security of network.

Intrusion Detection System has basically two types:

1. Without Signature Based Detection
2. With Signature Based Detection.

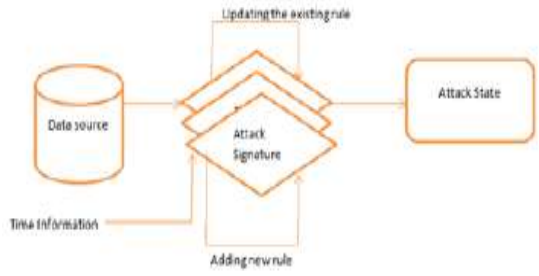

Figure 1: Without Signature Based Detection

Without Signature Based Detection is typically conceived as a more powerful method due to its theoretical potential for

addressing novel attacks. The second approach is most widely used and it detects only known attacks that have their signature included in the dataset. Many challenges need to be considered when building an intrusion detection model, such as obtaining a high attack Detection Rate without generating many false alarms. Early in the decade, researchers focused on using rule based expert systems and statistical approaches. But when encountering larger datasets, the result of rule-based expert systems and statistical approaches become worse. There are many data mining techniques have been introduced to solve the problem. Artificial neural network (ANN) is one of the most widely used data mining and has been successful in solving many complex practical problems due to encounter of large traffic data set. Based on a study of latest research literatures, there are quite a lot of research that attempts to relate data mining and machine learning techniques to the intrusion detection system so as to design more intelligent intrusion detection model.

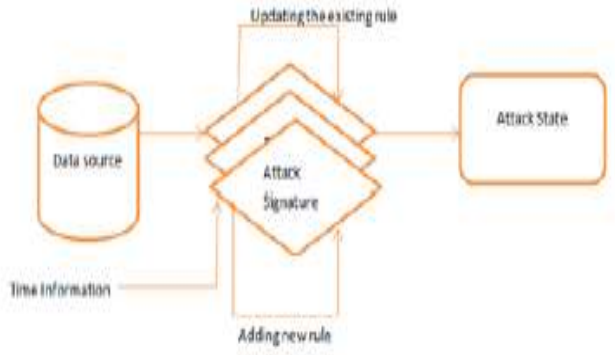

Figure 2: With Signature Based Detection

II. LITERATURE SURVEY

In this section, related literature about support vector machine approach and preparation of datasets for data mining activity will be reviewed and discussed.

Annie George [4], Anomaly detection has emerged as an important technique in many application areas mainly for network security. Anomaly detection based on support vector machine learning algorithms considered as the classification problem on the network data has given that. Dimensionality reduction and classification algorithms are explored and evaluated using KDD 1999 dataset for Intrusion Detection System. The result shows the decrease in execution time for the classification as they reduce the dimension of the input data and also the precision and recall parameter values of the classification algorithm shows that the Support Vector Machine with Principle Component Analysis method is more accurate as the number of misclassification decreases.

Kui W. Mok [5], there is produced the need to update an installed intrusion detection system (IDS) due to new attack methods or upgraded computing environments. This paper describes a data mining technique for adaptively building Intrusion Detection System (IDS) models. The central idea is to utilize auditing programs to extract an extensive set of features that describe each network connection or host session, and apply data mining programs to learn rules that accurately capture the behavior of intrusions and normal activities. These rules can then be used for signature based detection and novel detection. We discuss the idea of
our data mining programs, namely, classification, meta-learning, association rules, and frequent episodes. We analyze the results of applying this technique to the
extensively gathered network audit data for the 1998 DARPA Intrusion Detection Evaluation Program.

V. Jyothsna, V. V. Rama Prasad, K. Munivara Prasad [6], With the advent of anomaly-based intrusion detection systems, many approaches and techniques have been develop to track novel attacks on the systems. High detection rate of 98 percent at a low alarm rate of 1 percent can be achieved by using these techniques. Though
anomaly-based approaches are efficient, signature-based detection is preferred for mainstream implementation of intrusion detection systems. As a variety of anomaly
detection techniques were suggested, it is difficult to compare the strengths,Weaknesses of these methods. The reason why industries do not favor the anomaly based intrusion detection methods can be well understood by validating the efficiencies of the all the methods. To investigate this issue, the current state of the experiment practice in the field of anomaly-based intrusion detection is reviewed and survey recent studies in this. This paper contains summarization study and identification of the drawbacks of formerly surveyed works.

CHEN Bo, Ma Wu [7], the effective way of improving the efficiency of intrusion detection is to reduce the heavy data process workload. In this paper, the dimensionality reduction use of technology in the classic dimensionality reduction algorithm principal component to analysis large scale data source for reduced-made features of the original data be retained and improved the efficiency of intrusion
detection. And use BP neural network training the data after dimensionality reduction, will be effective in normal and abnormal data distinction, and achieved good results.

Paul Dokas , Vipin kumar [8], in which they gives an overview of our research in building rare class prediction models for identifying known intrusions and their variation and anomaly detection schemes for detecting novel attacks whose nature is unknown. Disadvantage of this paper is that due to the fact that the number of instances of U2R and R2L attacks in the training data set is very low, these numbers are not adequate as a standard performance measure. It could be biased if we use these numbers as a measure for performance of the system.

III METHODOLOGIS

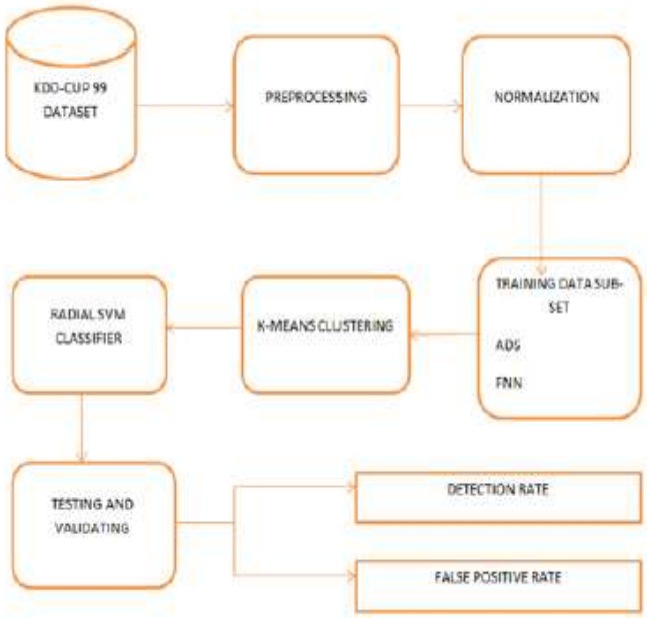

Figure 3: Architecture for Proposed IDS

The system architecture of proposed technique as shown in the figure 3 consists of 5 step methodology.

1. The input data set needed for experimentation is prepared by conducting relevance analysis on KDD Cup 1999 data set in order to reduce the irrelevant features which will not contribute for intrusion detection.
2. The input dataset is divided into two parts, Training Data set Testing Data Set. The Training data has a 4 attack data set and 1 normal dataset by using k- means clustering. Testing data set is used as a duplicate content can detect.

3. Fuzzy-neural Network (FNN) training is given to each of the attack data set and normal dataset, 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, D3, D4,. . . . . .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 ij which has membership value of each of the data.

5. Using Radial Support Vector Machine to detect intrusion has detected or not.

3.1 Data Collection

The KDD data set is the large amount of connection records. Analyzing KDD training and testing datasets, we will found that about 78 percent and 75 percent of the records are duplicated in the training and testing dataset, respectively. This large amount of redundant records in the training dataset will cause k-means algorithms to be used towards the more frequent records, and thus prevent it from machine learning unfrequented records which are usually more harmful to networks such as user to request (U2R) attacks. The existence of these repeated records in the testing dataset, on the other hand, will cause the experimental results to be biased by the methods which have better detection rates on the given records. This paper, it gives an overview of the data set used for intrusion detection. This data set contains 7 weeks of network traffic and 2 weeks of testing data. The raw data was about 4 gigabytes of compressed binary TCP dump data from the of network traffic generated. This was processed into about 5 million connection records, each of which is a vector of extracted feature values of that network connection. This data set of the 4,900,000 connection records was used as the data set for the 1999 KDD intrusion detection contest [11]. In particular, MIT Lincoln Labs DARPA 98 intrusion detection evaluation datasets have been employed to design and test intrusion detection systems. This formed the KDD 99 intrusion detection is a knowledge discovery dataset. The KDD 99 intrusion detection datasets are the advanced DARPA-98 initiative, which provides designers of intrusion detection systems (IDS) with a benchmark on which to evaluate different methodologies [3]. . Basically different 37 types of attack in the data set. The attacks identified exactly one of the four: User to Root (U2R); Remote to Local (R2L); Denial of Service (DOS); and Probe.

DOS: It has greater frequency and it can easily separate normal attack.
R2L: It has no any account on the victim machine, hence tries to gain access.
U2R: It is difficult to achieve detection accuracy. It has local access to victim machine.
Probe (Probing): It is same as DOS. Data preprocessing comprises following components including document conversion,
feature selection and feature weighting. The functionality of each component is described as follows:

1. Dataset prepared with DOS attack which include smurf, Neptune, back, teardrop and POD ping of death attacks anomaly.
2. Feature selection: reduces the dimensionality of the data space by removing irrelevant or less relevant feature selection criterion.
3. Document conversion: converts different types of documents such as gz, tcpdump to csv file and arff (Attribute-Relation File Format) data file format.
4. Totally we considered 11850 data points for our project.

3.2 K-Means Clustering

Support vector machine (SVM) has two types-
1. Supervised
2. Unsupervised
We use unsupervised learning machine that solve the clustering problem. The purpose of k-means algorithm is to classify a given KDD 99 data set into a certain number of clusters, i.e. Attack dataset or normal. The pseudo code for the adapted K-Means algorithm is presented as below

1. Choosing random k data points as initial Cluster Mean (Cluster center).
2. Again 1st step.
3. For each data point x from C.
4. Compute the distance x and takes its centroid.
5. Assign x to the nearest attack.
6. Finally done.
7. Re-compute the mean for current ADS or normal attack.
8. until reaching stable attack.
9. Use these centroid for normal and anomaly detection.
10. Calculate distance of centroid from normal and novel attack centroid points.
11. If distance(X, Cj)is less than equal to 5.
12. Then novel attack found; exit
13. Otherwise then
14. X is normal attack;

The k-means clustering algorithm is based on finding data clusters in a data set by keeping minimized cost function of dissimilarity measure. For each data point to be clustered, the cluster centroid with the minimal Euclidean distance from the data point will be the cluster for which the data point will be a member.

$$J = \sum_{J=1}^{K} \sum_{i=1}^{n} \| x_i^{(j)} - C_j \|^2$$

3.3 Fuzzy Neural Network

FNN is a biologically inspired form of distributed Computation. It can handle all kind of information. The connection between any two attribute has some weight, which is used to determine how much one attribute will affect the other attribute. A subset of the attribute acts as Input nodes and another subset acts as output nodes, which perform summation and threshold. The FNN has successfully been applied in different fields. The feed-forward neural network trained with the Back-propagation Neural Network (BPNN) algorithm is a common tool for intrusion detection system. FNN module aims to learn the pattern of every attack dataset and normal. It is simple processing units, and connections between them. In this paper, we will analyze classic feed forward neural networks trained with the Back-propagation Neural Network algorithm to predict intrusion detection. A feed-forward neural network has an input layer, an output layer, with one or more hidden layers in between the input and output layer. The Fuzzy Neural Network functions as follows: each node in the input layer has a signal xi th as network's input, multiplied by a weight.

Classification of the data point considering all its attributes is a very difficult task and takes much time for the processing, hence decreasing the number of attributes related with each of the given input i.e. attack dataset and normal dataset. Executing the reduced amount of training data also results in decrease of false positive rate and the improved performance of the classifier system. The main objective of this technique is to decrease the number of attributes associated with each data, so that classification can be mad in a simpler and easier way. Fuzzy-Neural network classifier is employed to efficiently decrease the number of attack.

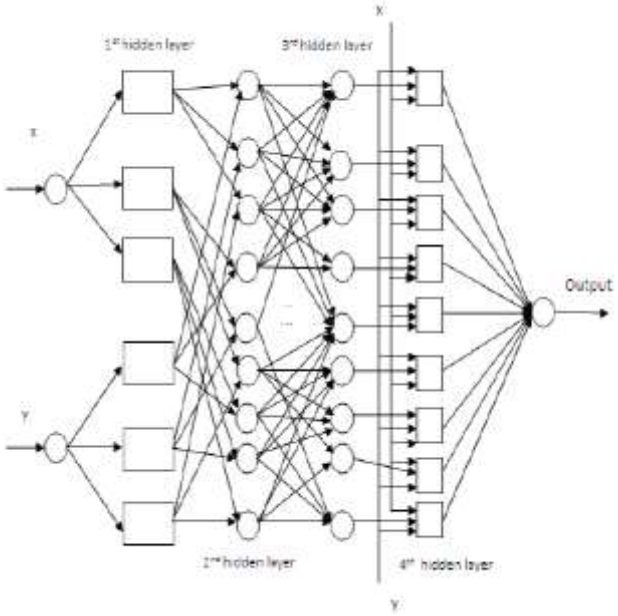

Figure 4: Neuro-fuzzy architecture

3.4 Radial SVM Classification

In our system, we are using radial SVM classifiers for the final classification. In the final classification, the data is binary classified to detect intrusion is happen or not. The attack dataset and normal dataset is trained with neuro-fuzzy after the initial clustering as we have discussed earlier, then the support vector is generated for reducing the attribute. Here in the process, each of the data is fed into each of the neural classifier to get the intrusion detection so the data value gets distorted and after passing through the K neuro fuzzy classifier. The vector array S= D1, D2, ...DN where, Di is the ith Data and N is a total number of input data. Here, after training through the neuro-fuzzy the attribute number reduces to K numbers. Di = a1,a2,a3,a4.....ak here the Di is the i th data governed by attribute values ai , where ai will have the value after passing through the i th neuro fuzzy. Total number of neuro-fuzzy classifiers trained will be K and Membership value Uij is defined by the equation below.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\| x_i - c_i \|}{\| x_i - c_k \|} \right)^{\frac{2}{m-1}}}$$

Hence, the Support vector is modified as S*=D*1, D*2,.....D*N where S* is the modified SVM vector which consists of modified data D*I, which consists of an extra attribute of membership value Uij There is input data which

had 37 attributes is now constrained to K+1. Use of radial Support vector machine results in obtaining better results from the classification process when compared to normal linear SVM. In linear SVM, the classification is made by use of linear hyper-planes whereas in radial SVM, nonlinear kernel functions are used and the resulting maximum-margin hyper-plane fits in a transformed feature space. The Gaussian Radial Basics function is given by the equation:

$$\phi(x - x_j) = \exp\left(-\frac{1}{2\,\delta_j^2} \| x - x_j \|^2\right) J = 1,2,\ldots N$$

Where j=1, 2,3,4,5,.... ...N. The jth input data point xj defines the center of radial basis function, the vector x is the pattern applied to the input. is a measure of width of jth Guassian function with center xj .

IV. DISCUSSIONS AND RESULT

Table.1 Data points taken for training and testing.

	Normal	DOS	PROBE	R2L	U2R
Training	12500	12500	1054	39	21
Testing	12500	12500	1054	38	21

Table. 2 Accuracy comparisons with existing method.

Different methods	PROBE	DOS	U2R	R2L
KDD CUP 99	83.3	97.1	13.2	8.4
Multi class SVM	75	96.8	5.3	4.2
Columbia Model	96.7	24.3	81.8	5.9
Decision tree	81.4	60.0	58.8	24.2
BPNN	99.3	98.1	89.7	48.2
Our technique	97.31	98.80	97.52	97.51

The first form of (figure 5) is Intrusion Detection System mining which is user is to open KDD-99 Dataset. When user is click on open KDD dataset button then KDD-99 Dataset is loaded. After that click on Perform k-means then clustered will be formed including (Id, Attribute, and Types).

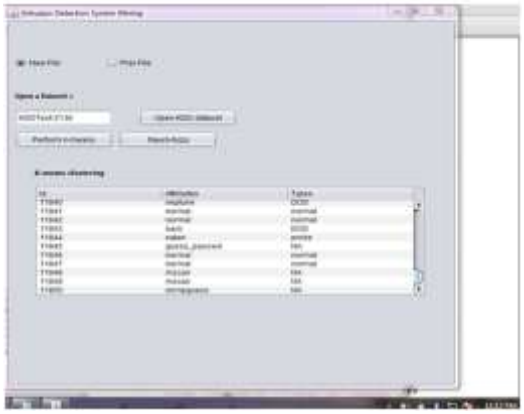

Figure 5: Screen shot for Intrusion Detection System mining form

This is the Fuzzy Neural Network (fig 6) form, it has four j-buttons (perform T-Norm, Fuzzy Logic, Radial SVM, and View.

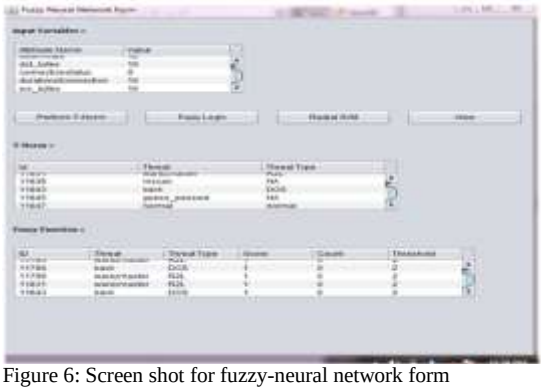

Figure 6: Screen shot for fuzzy-neural network form

At the last SVM form (figure 7) shows that Radial Support Vector Machine is to find intrusion is happen or not. The measurement used for evaluation of our proposed techniques are True positive (TP), False negative (FN), True negative (TN), and False positive (FP).

True Positive (TP) - A legitimate attack which triggers IDS to produce an alarm.

False Positive (FP) - An event signaling IDS to produce an alarm when no attack has taken place.

False Negative (FN) - A failure of IDS to detect present attack.

True Negative (TN) – It has no any alarms raised.

Table 3. Experimental result obtained for the training and testing dataset.

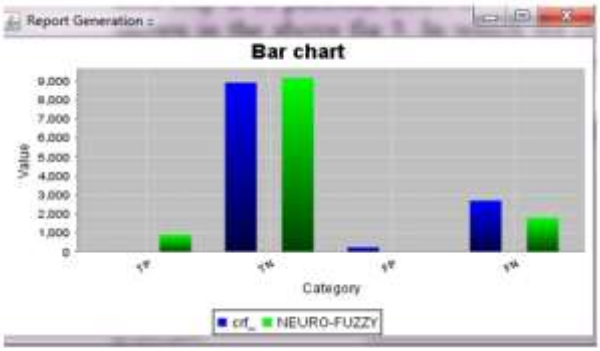

Figure 8: Performance comparison chart for SVM and Conditional Random Fields

V .CONCLUSIONS

The research on fuzzy neural network methods and support vector machine techniques to improve the detection rate by examining the behavior of the network as well as that of threats is done in the rapid force. The huge amount of dataset is increasing rapidly resulting in gradual rise in the security attacks. The current Intrusion Detection System is inappropriate, inaccurate and anomalous activity to update the audit data rapidly thus reduces the performances. This paper finding the architecture of the Intrusion Detection System along with advanced features of an intrusion detection system. In this paper we will analyzed the Fuzzy neural network approach and the Support vector machine approach in overcoming the challenges of the Intrusion Detection Technique. Further there is need to design the system which will overcome the current challenges of IDS and also the system must provide a high detection rate and false positive rate.

REFERENCES

[1] Adriana-Cristina Enache Intrusion Detection Based On Support Vector Machine Optimized With Swarm Intelligence, presented at 9th IEEE International on Applied Computational Intelligence and Informatics (2014), Timisoara, Romania.

[2] A.M.Chandrasekhar and K.Raghuveer, Intrusion Detection Technique by using K-means, Fuzzy Neural Network and SVM classifiers, presented at International Conference on Computer Communication and Informatics (ICCCI-2013), Coimbatore, INDIA.

[3] Sandip Ashok Shivarkar, and Mininath Raosaheb Bendre, Hybrid Approach for Intrusion Detection Using Conditional Random Fields, International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 3.

[4] Annie George, Anomaly Detection based on Machine Learning: Dimensionality Reduction using PCA and Classification using SVM, International Journal of Computer Applications (0975 to 8887) Volume 47 and No.21, June 2012.

[5] Kui W. Mok, A data mining framework for building intrusion detection model, In: Gong L., Reiter M.K. (eds.): Proceedings of the IEEE Symposium on Security and Privacy. Oakland, CA: IEEE Computer Society Press, pp.120 - 132, 1999.

[6] V. Jyothsna, V. V. Rama Prasad, K. Munivara Prasad, A Review of Anomaly based Intrusion Detection Systems, International Journal of Computer Applications (0975 to 8887) Volume 28 No.7, August 2011.

[7] CHEN Bo, Ma Wu, Research of Intrusion Detection based on Principal Components Analysis, Information Engineering Institute, Dalian University, China, Second International Conference on Information and Computing Science, 2009.

[8] Paul Dokas , Vipin kumar, Data Mining for Network Intrusion Detection, Proceeding of NGDM., pp.21-30, 2002.

[9] A.M.Chandrashekhar and K. Raghuveer. Performance evaluation of data clustering techniques using KDD Cup-99 Intrusion detection data set International Journal of Information Network Security (IJINS), Vol.1, No.4, pp.294-305, 2012.

[10] Jose Vieira, Fernando Morgado Dias, Alexandre Mota. Neuro-Fuzzy Systems: A Survey. Proceeding Internal Conference on Neural Networks and Applications, 2004.

[11] B. Pfahringer, Winning the KDD99 Classification Cup Bagged Boosting, SIGKDD Exploration, vol.1 pp. 65-66,2000.

[12] T. Ambwani, Multi class support vector machine implementation to intrusion detection, proceedings of IJCNN, PP.2300-2305, 2003.

[13] W. Lee, A framework for constructing features and models for intrusion detection system, information and system security, vol. 4 pp. 227-261,2000.

[14] J. H. Lee effective value of decision tree with KDD 99 intrusion detection dataset for intrusion detection system, proceding of 10th international conference on advanced communication technology Vol. 2, pp. 1170-1175, 2008.

