A survey of Network Intrusion Detection using soft computing Technique

Sumit Nigam[1] Ravindra Gupta[2] Gajendra Singh[3]
[1]M.Tech, [2]Asst. Prof., [3]H.O.D. CSE
[1,2,3]SSSIST, Sehore.

Abstract— with the impending era of internet, the network security has become the key foundation for lot of financial and business application. Intrusion detection is one of the looms to resolve the problem of network security. An Intrusion Detection System (IDS) is a program that analyses what happens or has happened during an execution and tries to find indications that the computer has been misused. Here we propose a new approach by utilizing neuro fuzzy and support vector machine with fuzzy genetic algorithm for higher rate of detection.

Keywords: neuro fuzzy, support vector machine, fuzzy genetic algorithm, Intrusion Detection System

I. INTRODUCTION

Internet has rapidly become one of the main communication methods in our society. Various types of internet application and usage are available more and more. Increasing usages of network applications also increase security risks to internet users. To prevent unwanted or dangerous threats, we have to be able to detect them first. Therefore, developing intrusion detection method is a challenging research issue because the network threats have different signatures and they evolve every day. The intrusion detection system at present must be able to detect new attacks. Intrusion detection system algorithms can be categorized into two types: supervised learning and unsupervised learning. A supervised learning is a technique that builds detection rule/model by learning pattern from provided information. The supervised learning normally has high detection rate and low false alarm rate. However, this technique can detect only known attacks. Therefore, it is not secure enough because in reality they are many new and unknown attacks in the internet. The second type of algorithms is an unsupervised learning technique. It is able to learn new/ unknown attacks without training information. However, it often has relatively lower detection rate and having high false alarm rate.

In 2007, Z. Banković et al. [1], proposed a Genetic algorithm for anomaly detection. They applied genetic algorithm as a rule based approach. They also proposed another method to pre-process dataset by using principal component analysis (PCA).

In 2008, T. P. Fries [3] proposed Fuzzy Genetic approach to network intrusion detection. This new adapted algorithm increased detection rate to 99.6 percent and only 0.2% of false positive.

In 2009, T. Komviriyavut et al. [4] proposed a real-time detection approach. They used packet sniffer to sniff network packets in every 2 seconds and preprocessed it into 12 features and used decision tree algorithm to classify the network data. The output can be categorized into 3 types which are DoS, Probe and normal. The result shows that this algorithm has 97.5 percent of detection rate. This technique is fast and able to use in real network. However, it was not designed to detect unknown attacks.

In 2011, Z. Muda et al. [5] proposed network detection solution by combining supervised learning technique and unsupervised learning technique. The KDD99 dataset was used to evaluate the performance of this algorithm. The detection rate was improved to 99.6 percent. However, this solution is not practical for real network because K-Means algorithm requires more time to process huge data in real networks which could lead to bottleneck problem and system clash.

In summary, previous research work discussed above did not pay enough attention to unknown intrusion detection. Some of them did consider the unknown attack, but they used the KDD99 dataset which is over 12 years-old. The network data is out of date and many recent attack types are not included in the dataset.

In this paper, we focus on network intrusion detection for unknown attack types meaning that the approach is able to detect new or unknown type of attacks in the network. In particular, the network intrusion detection system should be able to identify normal network activity and classify attack types. We are interested in designing an IDS technique using Fuzzy Genetic algorithm. The fuzzy rule is a supervised learning technique and genetic algorithm make fuzzy rule able to learn new attacks by itself. Moreover, this technique has high detection rate and robust. Therefore, we apply the fuzzy genetic algorithm approach to our online intrusion detection system i.e. the data is detected right after it arrived to the detection system. We evaluate our IDS in terms of detection speed, detection rate and false alarm rate.

II. PROPOSED METHODOLOGY

In this section we elaborated our new approach. First of all we present the whole work of the approach then we discuss the four modules i.e. K-Means clustering, neuro fuzzy training module, SVM Training Vector Module and radial classification module. The proposed intrusion detection technique initially clusters the given training data set by using k- means clustering into K- clusters where k is the number of clusters. In the next step, neuro fuzzy training used to train 'K' neural network where each of the data in particular clusters is trained with the respective neural network associated with each of the cluster. Subsequently vector for SVM is generated. This vector consist of attribute values obtained by passing each of the data through all of the trained neuro fuzzy classifier and an additional attribute which has membership value of each of the data. At the last step classification performed by using radial SVM to detect intrusion has happened or not. The block diagram of the proposed technique is given in fig. 1

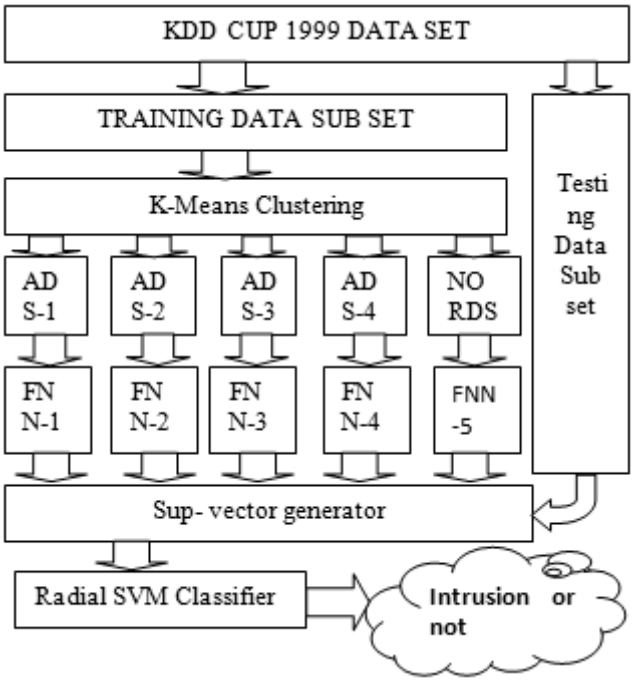

Fig. 1: Block diagram of proposed technique

The data set for our intrusion detection technique consists of large no of data, where each of the data considered has numerous attributes associated with it. Hence to perform classification considering all these attribute is a hectic and time consuming task. Processing and executing this limp amount of data results in increasing the error rate and also negatively affects the efficiency of the classifier system. In order to overcome the problem our proposed technique comes up with the solution where the number of attribute defining each of the data is reduced to a small number through a sequence of steps. This process ultimately result n making the intrusion detection more efficient and also yields a less complex system with a better result. Data set used to evaluate the validity of the proposed technique is prepared from KDD cup 1999 data set.

A. K- Means Clustering Module

The clustering algorithms are used to group unlabelled data. In our proposed technique, we are intended to group our input data set into different clusters based on types of intrusion. Since our input data set consist of normal data and different types of attacks, training data set grouped into 5 clusters using K-means clustering algorithm.

K-means is a prototype based, partitioned clustering technique that attempt to find a user specified number K of cluster, which are represented by centroid. The K-means algorithm processed the following

1) Define the number of cluster 'K'.
2) Initialize the k cluster centroids.
3) Iterate over all data points in the data set and compute the distances to the centroids of all the clusters.
4) Re calculate 'k' new centroids as per the centres of the clusters resulting from previous step.
5) Repeat step 3 until the centroids do not change any more.

The goal of clustering is typically expressed by an objective function that depends on the proximities of the points to one another or to the cluster centroids.

$$J=\sum_{J=1}^{K} \quad \sum_{i=1}^{n} \quad \square x_j^{(j)} - c_j \square^2$$

Where $\square x_j^{(j)} - c_j \square^2$ is a given chosen distance between a data point and cluster centre c_jis an indicator of the distance of the n data inputs from their respective cluster centres. Finally this algorithm aims at minimizing an objective function in this case a squared error function.

B. Neuro fuzzy training module

Neural networks are a significant tool for classification. The ability of high tolerance for learning by example makes neural network flexible and powerful in IDS. Neuro fuzzy refers to the combination of fuzzy set theory and neural network with advantages both. Neuro fuzzy incorporates fuzzy sets and linguistic model consisting of a set of IF-THEN rules. The main strength of neuro fuzzy systems is that they are universal approximators with the ability to solicit interpretable IF-THEN rules. The main advantages of neuro- fuzzy usage are that, it can handle any kind of information. It can manage imprecise, partial, vague or imperfect information. K-Means clustering results in the formation of 'K' cluster where each cluster will be type of intrusion or the normal data. Neuro fuzzy makes use of back propagation learning to find out the input membership function parameters and the least mean square method to find out the consequent parameter.

In the figure show the neuro fuzzy architecture. The first hidden layer maps the input variable correspondingly to each membership functions. In the second hidden layer, T-norm operator is used to compute the antecedents of the rules. The rules strength are normalised in the third hidden layer and subsequently in the fourth hidden layer. The output layer computes the result or output as the summation of all the signals that reach to this layer.

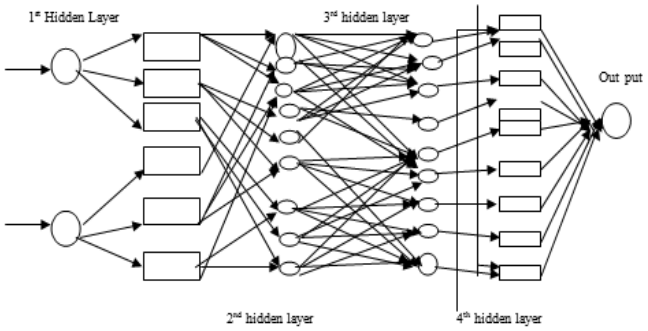

Fig. 2: Neuro- Fuzzy Architecture

C. SVM vector generation Module

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 data point is of paramount importance. The main purpose of the proposed technique is to decrease the number attributes associated with each data, so that classification can be made in a simpler and easier way.

The input data is trained with neuro- fuzzy after the initial clustering then the vector necessary for the SVM is generated. The vector arry 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.....ak}, here the Di is the ith data governed by passing the ith neuro- fuzzy. Total number of neuro fuzzy classifiers trained will be 'K', corresponding to 'K' clusters formed after clustering. Membership values μ_ij is defined as given by the equation below.

$$\mu_{ij} = \frac{1}{\sum_{k=i}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{2/m-1}}$$

Hence the training vector is modified as S* = {D*1, D*2.....D*N} where S* is the modified SVM vector which consist of modified data D*i, which consists of an extra attribute of membership value

μ_{ij}. D_1^*={a1,a2.....ak, μ_{ij}},hence the attribute number is reduced to K+1 where 'K' is the number of cluster formed in the start on.

D. Radial SVM classifier module

SVM classifier is used as it produces better results for binary classification when compared to the other classifiers. But use of linear SVM has the disadvantages of getting less accuracy result, over fitting results and robust to noise. These short coming are effectively suppressed by the use of the radial SVM where nonlinear kernel functions are used and the resulting feature space. In our proposed technique, nonlinear kernel functions are used and the resulting maximum margin hyper plane fits in a transformed feature space. When the kernel used is a Gaussian radial basis function, the corresponding feature space is a Hibert space infinite dimensions. The Gaussian radial basis function is given by the equation below.

$$c(x - x_j) = exp\left(-\frac{1}{2\alpha_j^2}\|x - x_j\|2\right)$$

J= 1,2,.....N

The 'j'th input data point xj defines the centre of radial basis function, the vector 'x' is the pattern applied to the input. αj is a measure of width of jth Gaussian function with centre xj.

The input dataset having large number of attributes is changed into that having K+1 attribute by performing the above step. The data with constrained number of attributes is given to the radial SVM, which is binary, classified to detect if there is any intrusion or not.

III. CONCLUSION

It is impossible to prevent security violation completely by using the existing security technologies. Accordingly intrusion detection is important component in network security. IDS helps the information security community by increasing detection efficiency, reducing the manpower needed in monitoring and helping to learn new vulnerabilities by providing legal evidence.

In this paper we present an effective technique for intrusion detection by making use of k- means clustering, fuzzy neural networks and radial support vector machine. We took the help of k means clustering technique to make large heterogeneous training data set in to a number of homogenous subsets. As a result complexity of each subset is reduced and consequently the detection performance is increased. After initial clustering in the proposed technique, training will be given to fuzzy neural network and later SVM will be used to perform final classification. We have used confusion matrix for the purpose of evaluation of our proposed technique.

REFERENCES

[1] David Wagner and Paolo Soto, "Mimicry Attacks on Host Based Intrusion Detection Systems" Proceedings of the 9th ACM conference on Computer and communications security, pp. 255 - 264, 2002.

[2] Ghanshyam Prasad Dubey, Neetesh Gupta and Rakesh K Bhujade, "A Novel Approach to Intrusion Detection System using Rough Set Theory and Incremental SVM", International Journal of Soft Computing and Engineering (IJSCE), vol.1, no.l, pp.1448, 2011.

[3] Hansung Lee, Jiyoung Song, and Daihee Park, "Intrusion Detection System Based on Multi-class SVM", Dept. of computer & Information Science, Korea Univ., Korea, pp. 51 U519, 2005.

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

[5] Horeis, T, "Intrusion detection with neural network - Combination of self-organizing maps and redial basis function networks for human expert integration", a Research report 2003. Available in hap://ieee-cis.org/Jiles/ EA C-Research-2003-Report-Horeis.pdf

[6] Han, S. J., & Cho, S. B. "Evolutionary neural networks for anomaly detection based on the behavior of a program", IEEE Transactions on Systems, Man and Cybernetics (Part B), 36(3), pp. 559-570, 2005.

[7] Chen, Y. H., Abraham, A., & Yang, B, "Hybrid flexible neural-tree-based intrusion detection systems", International Journal of Intelligent Systems(IJIS), 22(4), pp. 337-352, 2007.

[8] A.M.Chandrashekhar and K. Raguveer. "Performance evaluation of data clustering techniques using KDD cup 99 intrusion data set" International journal of information and network security(IJINS), Vol 1(4), pp. 294-305, 2012.

[9] R. Jang. "Neuro-Fuzzy Modeling: Architectures, Analysis and Applications", Ph D Thesis, University of California, Berkley, 1992.

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

[11] J.J. Buckley, Y. Hayashi and E. Czogala, "On the equivalence of neural nets and fuzzy expert systems, Fuzzy Sets & Systems", A Research Report, pp. 129-134. 1993.

[12] Aickelin, U, Twycross, J., Hesketh-Roberts, T "Rule generalization in intrusion detection systems using SNORT", International Journal of Electronic Security and Digital Forensics, 1(1), pp. 101-116, 2007.

[13] T. G. Dietterich and G. Bakiri. "Solving multiclass learning problems via error-correcting output codes",

Journal of Artificial Intelligence Research (JAIR) vol 2, pp. 263-286, 1995.

[14] M. Tavallaee, E. Bagheri, W. Lu and A. A. Ghorbani. "A detailed analysis of the KDD CUP 99 data set", Proceedings IEEE international conference on Computational intelligence for security and defense applications, pp. 53-58, Ottawa, Ontario, Canada, 2009.

[15] Dokas, P., Ertoz, L., Lazarevic, A., Srivastava, J., & Tan, P. N. "Data mining for network intrusion detection", Proceeding of NGDM., pp.2130, 2002.

[16] Wen Zhu, Nancy Zeng, Ning Wang. "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations", Proceedingsof NESUG Health Care and Life Sciences, Baltimore, Maryland, 2010.

[17] B. Pfahringer, "Winning the KDD99 Classification Cup: Bagged Boosting," SIGKDD Explorations, vol. 1, pp. 65-66, 2000.

[18] R. Agarwal and M. V. Joshi, "PNrule: A New Framework for Learning Classifier Models in Data Mining," in A Case-Study in Network Intrusion Detection, 2000.

[19] T. Ambwani, "Multi class support vector machine implementation to intrusion detection," Procedings ofUCNN, pp. 2300-2305, 2003.

[20] K. K. Gupta, B. Nath, and R. Kotagiri, "Layered Approach using Conditional Random Fields for Intrusion Detection," IEEE Transactions on Dependable and Secure Computing, vol. 5, 2008.

[21] W. Lee and S. Stolfo, "A Framework for Constructing Features and Models for Intrusion Detection Systems," Information and System Security, vol. 4, pp. 227-261, 2000.

[22] J.-H. Lee, J.-H. Lee, S.-G. Sohn, J.-H. Ryu, and T.-M. Chung, "Effective Value of Decision Tree with KDD 99 Intrusion Detection Datasets for Intrusion Detection System," Proceedings of 10th International Conference on Advanced Communication Technology, vol. 2, pp. 11701175,2008.

[23] Tich Phuoc Tran, Longbing Cao, Dat Tran and Cuong Due Nguyen, "Novel Intrusion Detection using Probabilistic Neural Network and Adaptive Boosting", International Journal of Computer Science and Information Security (IJCSI),Vol. 6, No. 1, pp.83-91, 2009.

