www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 11 November, 2014 Page No. 9124-9127

A Novel and Hybrid Technique for Efficient Intrusion Classification

Richa Shivhare[1], Sushil Chaturvedi[2]
Student M.Tech Asst Prof
[1] Department of Computer Science & Engineering
Shriram College of engineering & Management [SRCEM] Gwalior (MP), India
Richagupta1414@gmail.com
[2] Department of Computer Science & Engineering
Shriram College of engineering & Management [SRCEM] Gwalior (MP), India
sushilchaturvedi786@gmail.com

Abstract: ***The intrusion detection system (IDS) is one way of protecting a computer network. This kind of technology enables users of a network to be aware of the incoming threats from the Internet by observing and analyzing network traffic. The proposed technique involved four steps, first apply DBSCAN clustering which is used to make clusters, based on this obtained clusters we trained the network with by Back Propagation algorithm. We also apply Information Gain based Feature Selection method to identify the important features of the network. We trained the network once with all features and then reduced features this shows that we attain high detection rate and in efficient time. The developed network is used to identify the occurrence of various types of intrusions in the system. The performance of the proposed approach is tested using KDD Cup'99 data set available in the MIT Lincoln Labs. Simulation result shows that the proposed approach detects the intrusions with high detection rate and low false alarm and in high efficiency in terms of time.***

Keywords: Artificial Neural Network, supervised Learning Algorithm, Classification, DBSCAN Clustering information gain. interpretation.

1. Introduction

The importance of the security measures grows bigger as the protection of data (e.g., company data and research data) and the protection against computer related malicious code or intrusion attacks (virus, worm, Trojan horse etc.) are becoming more and more frequent for almost everyone working with computers. This kind of technology enables users of a network to be aware of the incoming threats from the Internet by observing and analyzing network traffic. During these processes the IDS will gather information from the network traffic, which will be used to determine whether the traffic holds suspicious content. Upon suspicious behavior in the network traffic, the IDS program can be set to warn its administrator either by mail or SMS, and in most cases write out to log files, which the administrator can read and discover possible intrusion. The IDS does not prevent an intrusion like a firewall which closes ports entirely. The IDS lets the traffic flow but sees the traffic and detects intrusion without really doing anything about it. The rest is up to the administrator or the security policy.

Recently, Artificial Neural Networks have been successfully applied for developing the IDS. ANN has the advantage of easier representation of nonlinear relationship between input and output and is inherent by fast. Even if the data
were incomplete or distorted, a neural network would be capable of analyzing the data from a network. Feature selection is found to be more relevant to Intrusion detection
system since the selected features retain their physical

2. Related Work

The Intrusion Detection studies till 2013 is provided in the paper [1] by [A.M Chandrasekhar] [k Raghuveer]. There are various categories of method to detect attack classes such as machine learning methods, statistical methods, etc. It is observed that much of the previous work used traditional statistical methods to bring out the results, but very few studies have used machine learning methods. Recently, the trend is shifting from traditional statistical methods to modern machine learning methods. In Last Decade, rule based expert system & Statistical approach detect some well known attack with high detection rate but difficult to detect new type of attack.

In paper [2][Duo Liu, Chung-Horng Lung, Ioannis Lambadaris] and [Nabil Seddigh] said that K-means clustering and Gaussian Mixture Model (GMM) are effective clustering techniques with many variations and easy to implement. Fuzzy clustering is more flexible than hard clustering and is practical for intrusion detection because of the natural treatment of data using fuzzy clustering. Fuzzy c-means clustering (FCM) is an iteratively optimal algorithm normally based on the least square method to partition data sets, which has high computational overhead.
In Paper [4] [Iftikhar Ahmad] and [Azween B Abdullah Saudi Arabia] in his paper, adopts a supervised neural network phenomenon that is majorly used for detecting security attacks. The proposed system takes into account Multiple Layered

Perceptron (MLP) architecture and resilient back propagation for its training and testing. The system uses sampled data from Kddcup99 dataset, an attack database that is a standard for evaluating the security detection mechanisms.
The developed system is applied to different probing attacks. Furthermore, its performance is compared to other neural networks' approaches and the results indicate that their approach is more precise and accurate in case of false positive, false negative and detection rate. In Paper[10] This paper investigates the application of the Feed Forward Neural Network trained by Back Propagation algorithm for intrusion detection. Mutual Information based Feature Selection method is used to identify the important features of the network. If a variable has high value of mutual information with respect to the output, then this variable must have significant effect on the output value which is to be estimated.

3. Proposed Model for Intrusion Detection

The proposed framework consists of three parts Feature reduction, clustering and classification to detect intrusion in network traffic dataset. The feature reduction method is used to utilize the time needed for evaluation of performance while clustering(DBScan) help us to find complicated cluster shapes with only one assigned input parameter in a very quickly manner while classification using Back propagation neural network learns by example. You give the algorithm examples of what you want the network to do and it changes the network's weights so that, when training is finished, it will give you the required output for a particular input.

The Study is aim to covering:

1 Data Collection using KDD Cup dataset
2 Apply Clustering DBSCAN
3. Trained Network with all features
4. Feature Reduction Information Gain method now trained with reduced features
5. Testing the network to detect normal data and attack data.
6. Measure the performance of system in terms of accuracy, precision, recall and f-measure of several attack categories and normal data.

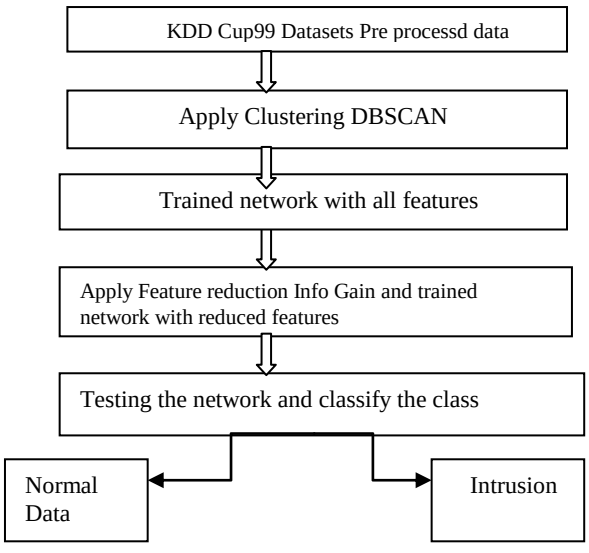

Figure1: Intrusion Detection Model

a) Data Collection

There are two ways to build IDS, one is to create our own simulation network, and collect relevant data and the other one is by using previously collected datasets. Some of the popularly used IDS datasets [14] are DARPA 1998 data set, DARPA 1999 data set and KDD Cup 1999 data set which are available in the MIT Lincoln Labs. In this work, we use KDD Cup 1999 data set for developing the IDS.

b) Data Preprocessing and Attribute Selection

Before training the neural network, the dataset should be preprocessed to remove the redundancy present in the data and the non-numerical attributes should be represented in numerical form suitably. Attribute selection methods information gain. Based on the entropy of a feature, information gain measures the relevance of a given feature. If the feature is relevant, in other words highly useful for an accurate determination, calculated entropies will be close to 0 and the information gain will be close to 1.

c) DBSCAN Clustering

The DBSCAN algorithm can identify clusters in large spatial data sets by looking at the local density of database elements, using only one input parameter. Furthermore, the user gets a suggestion on which parameter value that would be suitable. Therefore, minimal knowledge of the domain is required. The DBSCAN can also determine what information should be classified as noise or outliers. In spite of this, its working process is quick and scales very well with the size of the database almost linearly.

d) BPN Algorithm

- Before the BP can be used, it requires target patterns or signals as it a supervised learning algorithm.
- Training patterns are obtained from the samples of the types of inputs to be given to the multilayer neural network
- The purpose is to minimize the error between the target and actual output and to find Δw.
- The error is calculated at every iteration and is back propagated through the layers of the ANN to adapt the weights.
- The weights are adapted such that the error is minimized.
- Once the error has reached a acceptable minimum value, the training is stopped, and the neural network is reconfigured in the recall mode to solve the task

Bias weights are used with bias signals of 1 for hidden (j) and output layer (k) neurons.
In many ANN models, bias weights (θ) with bias signals of 1 are used to speed up the convergence process. The learning parameter is given by the symbol η and is usually fixed a value between 0 and 1, however, in many applications nowadays an adaptive η is used. Usually η is set large in the initial stage of learning and reduced to a small value at the final stage of learning. A momentum term α is also used in the G.D.R. to avoid local minimas

e) Feature *Selection Method*

Feature selection improves classification by searching for the subset of features, which best classifies the training data. Feature selection leads to savings in measurement cost and the selected features retain their original physical interpretation. Hence, feature selection is more relevant to Intrusion Detection System. Based on the entropy of a feature, information gain measures the relevance of a given feature, in other words its role in determining the class label. If the feature is relevant, in other words highly useful for an accurate determination, calculated entropies will be close to 0 and the information gain will be close to 1. Once the information gain value of feature variables is evaluated, the variables are ranked, with the variable having the high information gain value at the top and so on.

Information *Gain*

Let S be a set of training set samples with their corresponding labels. Suppose there are *m* classes and the training set contains *si* samples of class *I* and s is the total number of samples in the training set. Expected information needed to classify a given sample is calculated by:

$I(s_1, s_2 \ldots\ldots s_m) = -\sum_{j=1}^{m} \frac{si}{s} log2 \frac{si}{s}$

A feature *F* with values *{ f1, f2, …, fv }* can divide the training set into v subsets *{ S1, S2, …, Sv }* where *Sj* is the subset which has the value *fj* for feature *F*. Furthermore let *Sj* contain *sij* samples of class *i*. Entropy of the feature *F* is

$E(F) = \sum_{j=1}^{vm} \frac{s1j+\cdots\ldots smj}{s} \times I(S1j \ldots \ldots Smj) \frac{si}{s}$

Information gain for F can be calculated as:

$Gain\ (F) = I(S_1,\ldots,S_m) - E(F)$

4. Experimental Results

The original data contain 744MB data with 4,940,000 records. A ten percent subset of this data contain 75MB with 4,94,021 records which approximately contain 20% normal patterns and the rest 80% of patterns are with attacks belonging to four categories (DOS, Probe, U2R and R2L). Among them we have selected 10,000 records randomly for developing the Neural Network. The details of the records selected for training and testing the Neural Network is given in Table 1.

Table 1: Total number of Samples 20000

Data distribution	*Normal*	*Dos*	*Probe*	*U2R*	*R2L*
Training	3000	3000	2377	39	3000
Testing	5000	5000	4377	78	5000

Among the 41 input features, 32 features are continuous variables and 9 features are discrete variables. Suitable integer numbers are assigned to these discrete variables. For example, for the discrete variable protocol type which describes the type of the protocol we have assigned 1 for tcp, 2 for udp, 3 for http and so on. Accordingly suitable integer numbers are assigned to other discrete variables also. The output attack label is represented as [1, 0 0 0 0] for Normal, [0 1 0 0 0] for DOS, [0 0 1 0 0] for Probe, [0 0 0 1 0] for R2L and [0 0 0 0 1] for U2R. Two different ANN models were developed for intrusion detection, one with all features and another with reduced features. The network is trained with least means square algorithm until it reaches the mean square error of 0.01.

After training the performance of the network model is evaluated with 19,455 test data. The trained Neural Network classified 19,455 data correctly. Once trained the network with all features which shows an overall detection rate of overall accuracy 97.54% and time taken is 0.56 sec. after that test with reduced features and get 98.46 % overall accuracy and testing time taken is 0.32 sec.

The neural network model is developed using Net Beans IDE 7.4 with 1.90 GHz processor with 2 GB of RAM. The details of the model developed are given below:

Testing with ANN (Artificial neural network) and RNN (Recurrent neural network)

Case (I): Testing with ANN with reduced features (32 selected)

In this case, apply DBSCAN clustering trained the network by applying back propagation neural network with features 32 and testing is done with 5000 data in each class, achieve the overall accuracy 97.94 % and time is 0.56 sec.

Table 2: Testing with Artificial Neural Network with reduced features

Classes of Data	Calculated % wise				Overall Accuracy	Total tim taken in testing(se
Types	**Accuracy**	**Precision**	**Recall**	**F-Measure**		
NORMAL	96.47	99.31	88.74	93.73	97.94%	0.56 seconds
DoS	99.98	99.94	100	99.97		
PROBE	96.49	80.67	98.86	88.85		
U2R	99.54	96.77	76.92	85.71		
R2L	99.92	99.84	99.86	99.85		

Case (II): Testing with RNN with reduced features (32)

In this case, testing the network with RNN with features 32 and testing is done with 5000 data in each class, achieve the overall accuracy 98.68 % and time is 0.32 sec.

Table 3: Testing with Recurrent Neural Network with reduced features

Classes of Data	Calculated % wise				Overall Accuracy	Total time taken in testing(sec)
Types	Accuracy	Precision	Recall	F-Measure		
NORMAL	94.88	85.50	99.68	92.05	98.56 %	0.32 seconds
DoS	99.98	99.54	100.00	99.97		
PROBE	94.97	99.36	64.87	78.49		
U2R	99.95	87.50	89.74	88.61		
R2L	99.90	99.91	99.73	99.82		

5 Conclusion

Comparison of results shows that our hybrid model gives better performance. Reduced feature set plays an important role in intrusion detection model. This reduced feature set improves the performance of learning algorithm and also reduces the computational cost. This further reduces the complexity of the classifier also. From the results the hybrid model better detects the attack and normal class data. We compared different clustering and classification techniques for network traffic anomaly detection using Resilient BPN, Gradient Descent and Genetic algorithm. More similar type of studies can be carried out on different traffic datasets to give generalized results across different organizations. We plan to repeat our study on larger datasets and some other traffic dataset which could be available in near future. In future studies, we will take into account some more categories of attack other than (Dos, u2r, r2l & probe) to get more accurate and efficient results.

6 References

[1] A.M Chandrasekhar, K. Raghuveer 2013 IEEE: "Intrusion *Detection Techniques by using k-means clustering, fuzzy neural network and SVM classifiers"* International Conference on Computer Communication and Informatics (ICCCI-2013) 978-1-4673-2907-1/13 Coimbatore India.

[2] Duo Liu, Chung-Horng Lung, Ioannis Lambadaris and Nabil Seddigh IEEE 2013: "*Network traffic anomaly detection using clustering techniques and performance*" Canadian conference of electrical and computer engineering (ccece) 978-1-4799-0033-6/13

[3] Vladimir Bukhtoyarov and Eugene Semenkin :" *Neural Networks Ensemble Approach for Detecting Attacks in Computer Networks"* WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia.

[4] Iftikhar Ahmad Azween B Abdullah Abdullah S Alghamdi:*"Application of Artificial Neural Network in Detection of Probing Attacks*" 2009 IEEE Symposium on Industrial Electronics and Applications (ISIEA 2009), October 4-6, 2009, Kuala Lumpur, Malaysia

[5] E.Kesavulu Reddy, Member IAENG: *"Neural Networks for Intrusion Detection and Its Applications"* Proceedings of the World Congress on Engineering 2013 Vol II, WCE 2013, July 3 - 5, 2013, London, U.K.

[6] Iftikhar Ahmad, M.A Ansari, Sajjad Mohsin. :"*Performance Comparison between Backpropagation Algorithms Applied to Intrusion Detection in Computer Network"* 9th WSEAS International Conference on Neural Networks (NN'08), Sofia, Bulgaria, May 2-4, 2008 ISBN: 978-960-6766-56-5.

[7] Shi-Jinn Horng Ming-Yang Su Yuan-Hsin Chen Tzong-Wann Kao: "*A novel intrusion detection system based on hierarchical clustering and support vector machines"*0957-4174/ 2010 Elsevier Ltd.

[8] Dr. Saurabh Mukherjee, Neelam Sharma*:"Intrusion Detection using Naive Bayes Classifier with Feature Reduction"* © 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of C3IT

[9] H. Günes Kayacık, A. Nur Zincir-Heywood, Malcolm I. Heywood: *"Selecting Features for Intrusion Detection:A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets"*

[10] P. Ganesh kumar and D.Devaraj*: "Intrusion Detection using Artificial Neural Network with Reduced Input Features"* ICTACT Journal on Soft Computing, July 2010, Issue: 01

[11] Shaik Akbar Dr.K.Nageswara Rao Dr.J.A.Chandulal: *"Intrusion Detection System Methodologies Based on Data Analysis"* International Journal of Computer Applications (0975 – 8887)Volume 5– No.2, August 2010

[12] DARPA Intrusion Detection Evaluation – MIT Lincoln Laboratory – (http://www.ll.mit.edu/IST/ideval)

[13] A. H. Sung, S. Mukkamala, 2003,: *"Identifying important features for intrusion detection using support vector machines and neural networks,"* in Proceedings of International Symposium on Applications and the Internet (SAINT 2003), pp. 209-17.

[14] KDD-cup dataset, http://kdd.ics.uci.edu/ databases/ kddcup99/ kddcup99.htm.

[15] Parveen Kumar, Nitin Gupta:" A *Hybrid Intrusion Detection System Using Genetic-Neural Network"* International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 National Conference on Advances in Engineering and Technology (AET- 29th March 2014)

[16] A. Zhong and C.F. Jia. 2004*, "Study on the applications of hidden Markov models to computer intrusion detection,"* in *Proceedings of the Fifth World Congress on Intelligent Control and Automation WCICA*, Vol. 5, pp. 4352-4356.

[17] M.Analoui, A.Mizaei, and P.Kabiri. 2005*, "Intrusion detection using multivariate analysis of variance algorithms,"* in Third International Conference on Systems ,Signals & Devices SSD05, Vol. 3.