Intrusion Detection System based on Fuzzy C Means Clustering and Probabilistic Neural Network

Rachnakulhare
M.Tech VI sem,
BUIT,BU Bhopal

Divakar Singh
Head of CSE deptt.
BUIT, BU Bhopal

ABSTRACT

Security is always an important issue especially in the case of computer network which is used to transfer personal/confidential information's, ecommerce and media sharing. Since the network is closely related to operating its conditions hence a careful observation & analysis of network characteristics could describe the state of the network such as network is under specific attack or operating normally. This paper presents an intrusion detection system based on fuzzy C-means clustering and probabilistic neural network which not only reduces the training time but also increases the detection accuracy. The proposed system is tested using KDD99 dataset and the simulation results shows that by selecting effective characteristics and proper training the detection accuracy rate up to 99% is achievable.

Keywords

Network Security, Neural Network, Intrusion Detection System, Fuzzy C-Means Clustering, KDD99 dataset.

1. INTRODUCTION

The Computer Network is designed for easier establishment on network with many facilities but making design like this also exposed it to network attackers and makes it a soft target for intruder hence extra care is required to be taken. To overcome these problems many techniques are proposed one most common is the modification in the protocol (since initially it is designed by considering the performance). Generally the modification in the protocol is toworks for specific attacks only and it may affect the performance also. Another problem with a Protocol modification is all nodes operating in that network must have same protocol or modifications must be compatible with standard one. Another approach which can work independently on any specific node or even on a separate observer unit is generally known as Intrusion Detection System (IDS). An IDS is a device or software application that monitors network or system activities for malicious activities or policy violations and produces reports to a Management Station. Although the IDS do not counter the attack but it can generate alarm. The analysis shows that both systems have their own limitations but a better system can be designed by combining the both algorithms the IDS system can used to initiates the specific modification in protocol and this could be the complete solution for the Intrusion Detection and Prevention System (IDPS).

2. RELETED WORK

Because of the importance of the subject already some work has been done some of them which are found most related and useful for making this paper is presented here. Bing Wu et.al. [1] presented great literature on the MANET attacks. Their work gives the detailed explanation of different attacks their behavior and their effect on network characteristics they also explained the security mechanism for some attacks although no simulation and mathematical details are provided. Another text on same topic is presented by Abhay Kumar Rai et.al. [2] the simulation and modeling of the different attacks on MANET using network simulator is explained in KarimKonate et.al. [3] the paper also discussed the protocols and their immunities to different attacks with analytical modeling and mathematical formulation. A graph based approach is proposed by Zhou Mingqiang et al [11] they proposed graph-based intrusion detection algorithm by using outlier detection method that based on local deviation coefficient (LDCGB). Compared to other intrusion detection algorithm of clustering, this algorithm is unnecessary to initial cluster number. Meanwhile, it is robust in the outlier's affection and able to detect any shape of cluster rather that the circle one only. Moreover, it still has stable rate of detection on unknown or muted attacks. Farah Jemili et.al. [4] presented the intrusion detection system based on Bayesian Network (BN). The BN is used to build automatic intrusion detection system based on signature recognition. The goal is to recognize signatures of known attacks, match the observed behavior with those known signatures, and signal intrusion when there is a match. Improved Support Vector Machine (SVM) based IDS model is presented in Jingbo Yuan et.al. [5] the paper discussed the method for improvement of SVM to achieve the higher accuracy. A data preprocessing and removal of similar data to reduce the training data size using k means clustering presented in [6][12] which shows significant improvement in training time with maintaining accuracy. One important requirement of classification is parameter selection because some of the features may be redundant or with a little contribution to the detection process. Gholam Reza Zargar and PeymanKabiri [7] investigate selection of effective network parameters for detecting network intrusions. The study shows that the major difficulty in develop the system like presented in [5][6][7] is that intrusions signatures changes broadly hence a large training dataset, parameter selection, data filtering and optimal classification is required. Besides mentioned limitation it has a great advantage of better classification without affecting the network performance.

3. KDD99 DATASET

This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 the Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment [9].The KDD99 dataset contains of seven weeks data and the following classes are used for the labeling of each data vectors.

• Denial of Service (dos): Attacker tries to prevent legitimate users from using a service.

• Remote to Local (r2l): Attacker does not have an account on the victim machine, hence tries to gain access.

• User to Root (u2r): Attacker has local access to the victim machine and tries to gain super user privileges.

• Probe: Attacker tries to gain information about the target host.

Subcategories and their Main categories are shown in table 1.

Table 1.List of Intrusions listed in KDD99 dataset.

Sub-Category	Main-Category
back	dos
buffer_overflow	u2r
ftp_write	r2l
guess_passwd	r2l
imap	r2l
ipsweep	probe
land	dos
loadmodule	u2r
multihop	r2l
neptune	dos
nmap	probe
perl	u2r
phf	r2l
pod	dos
portsweep	probe
rootkit	u2r
satan	probe
smurf	dos
spy	r2l
teardrop	dos
warezclient	r2l
warezmaster	r2l

The each data vector of the dataset contains 41 features which can be categorized into four categories:

• Basic Features: Basic features can be derived from packet headers without inspecting the payload.

• Content Features: Domain knowledge is used to assess the payload of the original TCP packets.

• Time-based Traffic Features: These features are designed to capture properties that mature over a 2 second temporal window.

• Host-based Traffic Features: Utilize a historical window estimated over the number of connections in this case 100 instead of time.

The detail descriptions of all features are available in [7].

4. FUZZY C-MEANS CLUSTERING

The Fuzzy C-Means (FCM) algorithm is one of the most widely used methods in fuzzy clustering. It is based on the concept of fuzzy c-partition, introduced by Ruspini (1970), Dunn (1974) and Bezdek (1981). In fuzzy clustering each data point belongs to every cluster by some membership value and the process of grouping is iterated till the change in the membership values of each data point stops changing. In many situations, fuzzy clustering is more natural than hard clustering. The detail working of Fuzzy C-Means clustering is explained by S. Nascimento et al [10].

5. PROBABILISTIC NEURAL NETWORK (PNN)

In a PNN, the operations are organized into a multilayered feedforward network with four layers, shown in fig 1. The input nodes are the set of measurements. The second layer consists of the Gaussian functions formed using the given set of data points as centers. The third layer performs an average operation of the outputs from the second layer for each class. The fourth layer performs a vote, selecting the largest value. The associated class label is then determined [8].

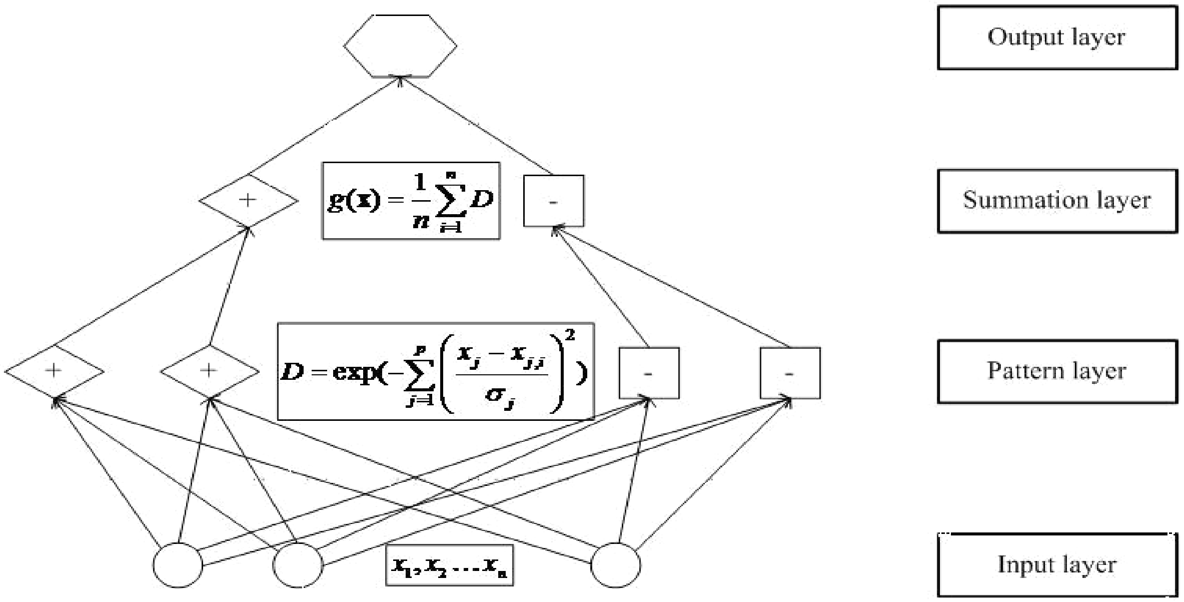
Output layer
Summation layer
Pattern layer
Input layer
+
+
+
-
-
-
$g(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} D$
$D = \exp(-\sum_{j=1}^{p}\left(\frac{x_j - x_{j,i}}{\sigma_j}\right)^2)$
$x_1, x_2 \ldots x_n$

Fig 1: Structure of PNN [9]

6. PROPOSED ALGORITHM

Inthe proposed system we firstly KDD99 dataset is used to generate the training vectors. Now after collecting all these parameters fuzzy C-means clustering is applied and the data with closer distances are eliminated then this data is used to train the neural network which is later used for detection of Intrusion. The algorithm can be described in detail by following steps:

Step 1: Read the given numbers of samples from KDD99 dataset.

Step 2: Filter selected features from the dataset for further processing.

Step 3: Partition the data into training and testing sets.

Step 4: Cluster the training dataset using the Fuzzy C-means Clustering.

Step 5: Select the fraction of data points from each edge of each cluster.

Step 6: Now train the Probabilistic Neural Network (PNN) using these vectors (data points) with their classification group.

Step 7: Test the trained PNN by the testing dataset.

Step 8: Calculate the performance of the trained system.

7. BLOCK DIAGRAM OFTHE PROPOSED SYSTEM

In figure 2 combined frame work the detection system is developed with three modules, namely, pre–processing phase, learning phase and testing phase. Above block diagram of the proposed system can be defined as follows.

• KDD99 dataset: Read the given numbers of samples from KDD99 dataset.

• Pre-processing: In pre-processing module, is to convert the data which is suitable for unsupervised learning by removing the labels from the dataset. Filter selected features from the dataset for further processing.

• Data partitioning: preprocessing data are used to partition into training & testing sets.

• Fuzzy c- means clustering: Fuzzy clustering plays an important role in solving problems in the areas of pattern recognition and fuzzy model identification. A variety of fuzzy clustering methods have been proposed and most of them are based upon distance criteria. One widely used algorithm is the fuzzy c-means (FCM) algorithm. Cluster the training dataset using the Fuzzy C-means Clustering. Select the fraction of data points from each edge of each cluster.

• Probabilistic neural network: Now train the Probabilistic Neural Network (PNN) using these vectors (data points) with their classification group. PNNs are faster to train and approach the Bayes optimal as the training set increases.

• Testing data: Test the trained PNN by the testing dataset.

• Performance calculation: Calculate the performance of the trained system.

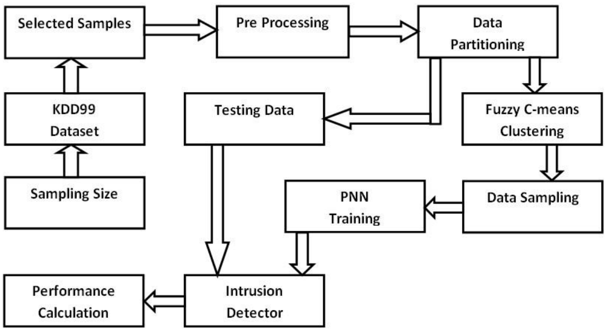

Figure 2: Block diagram of the proposed system.

8. SIMULATION RESULTS

The Simulation of the proposed work is performed using MATLAB 7.5 Neural network toolbox in IBM P4 dual core 2.4 Ghz processor with 2 GB of RAM and windows XP operating system. The results from this simulation are shows in table 2, 2.1 and 3, 3.1.

Table 2.Performance of PNN based System.

Dataset	TPR	TNR	FPR	FNR
1000	0.9021	0.7934	0.2066	0.0979
2000	0.7832	0.9061	0.0939	0.2168
3000	0.6837	0.9885	0.0115	0.3163
4000	0.7846	0.9035	0.0965	0.2154
5000	0.9021	0.7934	0.2066	0.0979

Table 2.1Performance of PNN based System.

Dataset	Acc.	Prec.	Recall	F-meas
1000	0.8050	0.3417	0.9021	0.4956
2000	0.8102	0.9674	0.7832	0.8656
3000	0.9540	0.8836	0.6837	0.7709
4000	0.8259	0.8914	0.7846	0.8156
5000	0.9167	0.7167	0.0833	0.0833

Table 3. Performance of Proposed System (Fuzzy C-means and PNN)

Dataset	TPR	TNR	FPR	FNR
1000	0.9680	0.9926	0.0074	0.032
2000	0.9931	0.9754	0.0246	0.0069
3000	0.9594	0.9984	0.0016	0.0406
4000	0.9866	0.9798	0.0202	0.0134
5000	0.9735	0.9888	0.0112	0.0265

Table 3.1 Performance of Proposed System (Fuzzy C-means and PNN)

Dataset	Acc.	Prec.	Recall	F-meas
1000	0.9900	0.9397	0.9680	0.9536
2000	0.9892	0.9931	0.9931	0.9931
3000	0.9940	0.9873	0.9594	0.9731
4000	0.9898	0.9867	0.9866	0.9866
5000	0.9911	0.9734	0.9735	0.9733

Simulated results of Detection Accuracy:

Figure 3. Show the graph is plotted between total no. of samplesV/S accuracy. Which show the comparisons between PNN only & PNN with clustering Detection accuracy? Blue line show PNN only & Red lines show PNN with fuzzy c-means clustering. By these comparisons Detection Accuracy performance of PNN with clustering much better than PNN only.

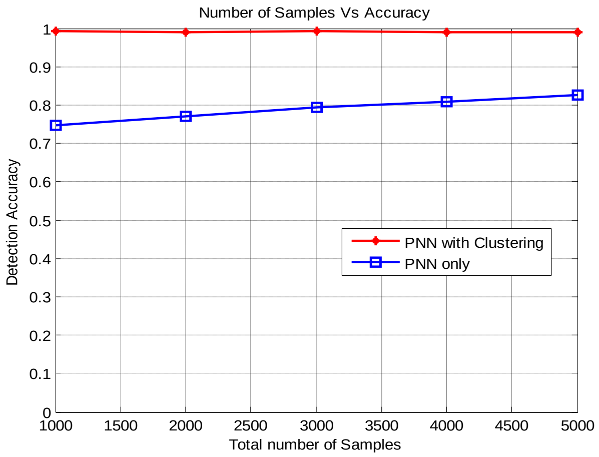

Figure3: Result analysis of PNN with clustering

& PNN only (Detection Accuracy)

9. CONCLUSION

A simulated resultthe model of the Intrusion detector is presented in this paper is not only capable of attack situation but can also classifying the individual attacks. The Detection accuracy of the system is up to 99% which is excellent also the algorithm have very low FPR (max 8.3%) hence reduces the chances of false alarming. The results also shows that it takes only 0.0075 seconds to identify the intrusion hence fast enough to prevent any loss due to delayed action. Further it could achieve much better performance by increasing the number of samples taken and increasing the number of characteristics parameter selected.

10. REFERENCES

[1] Bing Wu, Jianmin Chen, Jie Wu and MihaelaCardei,“A Survey of Attacks and Countermeasures in Mobile Ad Hoc Networks”, wireless/mobile network security,2006 Springer.

AUTHOR’ BIOGRAPHY

1. Ms. Rachnakulhare,presently she is a M.tech student of Computer Science & Engineering, BUIT, Barkatullah University, Bhopal. She Received degree B.E. from UIT RGPV, Bhopal, & 5 years teaching experience in UG respectively.

2. Dr. Divakar Singh, presently he is Head in Computer Science & Engineering, BUIT, Barkatullah University, Bhopal. He received degree B.E., form Barakatullah University Bhopal, M.TECH from RGTU, Bhopal & Ph.D. in Computer Science & Engineering. He has research interests in image analysis, image mining, soft computing & machine learning. He has 12 years of teaching experience in UG and 8 years of PG respectively, also guided more then 35 Mtech Dissertation

[2] Abhay Kumar Rai, Rajiv RanjanTewari and Saurabh Kant Upadhyay,“Different Types of Attacks on Integrated,MANET,InternetCommunication”,Internatioal Journal of Computer Science and Security (IJCSS),Aug. 2010.

[3] KarimKonate and Gaye Abdourahime “Attacks Analysis in mobile ad hoc networks: Modeling and Simulation”, Second International Conference on Intelligent Systems, Modelling and Simulation, 2011 IEEE.

[4] Farah Jemili, MontaceurZaghdoud and Mohamed Ben Ahmed,“A Framework for an Adaptive Intrusion Detection System using Bayesian Network”, 2007 IEEE.

[5] Jingbo Yuan, Haixiao Li, Shunli Ding and Limin Cao “Intrusion Detection Model based on Improved Support Vector Machine”, Third International Symposium on Intelligent Information Technology and Security Informatics, 2010 IEEE.

[6] Z. Muda, W. Yassin, M.N. Sulaiman and N.I.Udzir,“Intrusion Detection based on K-Means Clustering and OneR Classification”, 2011 IEEE.

[7] http://link.springer.com/chapter/10.1007%2F978-3-642-14400-4_50?LI=true#.

[8] http://www.personal.reading.ac.uk/~sis01xh/teaching/CY2D2/Pattern3.pdf

[9] http://voyagememoirs.com/pharmine/2008/06/22/probabilistic-neural-network-pnn/

[10] S. Nascimento, B. Mirkin and F. MouraPires “A Fuzzy Clustering Model of Data and Fuzzy c-Means”, Fuzzy Systems, the Ninth IEEE International Conference on May,2000.

[11] Zhou Mingqiang and Huang Hui, Wang Qian,“A Graph-based Clustering Algorithm for Anomaly Intrusion Detection”, The 7th International Conference on Computer Science & Education , July 2012. Melbourne, Australia.

[12] Sanjay Kumar Sharmai, PankajPande, Susheel Kumar Tiwari and Mahendra Singh Sisodia “An Improved Network Intrusion Detection Technique based on k-Means Clustering via NaIve Bayes Classificatio”, IEEE-International Conference On Advances In Engineering, Science and Management March , 2012.