Finding Anomaly With Fuzzy C-means ANN Using Semi-Supervised Approach

Ms Gadekar Supriya S.
Department of Computer Engineering of JSCOE
Handewadi Road, Hadapsar, India
Supriya.gadekar5@gmail.com

Prof.Shinde Sharmila M.
Department of Computer Engineering of JSCOE
Handewadi Road, Hadapsar, India
sharmi_anant@yahoo.co.uk

Abstract
The FC-ANN (Artificial Neural Network) is used to speed up the technique. The anomaly Outlier detection is primary in various data-mining applications. Outlier detection methods have been suggested for number of application such as, fraud detection, voting irregularity analysis, data cleansing, clinical trials, network intrusion, severe weather prediction, geographic information system, credit cards, athlete performance analysis and other data mining tasks proposed algorithm. This proposed system attaches the rough set theory, fuzzy set theory and semi-supervised learning to detect outliers as well as is a new try in area of outlier detection for semi-supervised learning. Without considering those points located in lower approximation of a cluster, proposed algorithm need to discuss the possibility of the points in boundary to be assigned as outliers and has number of advantages over semi-supervised outlier detection. In this proposed algorithm will be applied to various outlier detection fields which has only partially labeled samples, especially that does not make a certain judgment in uncertain conditions. The proposed system proposes the technique FC-ANN that may add parameters to speed up the technique.

Keywords: outlier detection; Semi-supervised learning ; fuzzy set ; C-means clustering.

Introduction
In number of application like criminal activity monitoring, fraud detection, network intrusion detection ,these deviated cases are more interesting as well as more useful than normal cases. Finding out such type of outlier can show the system faults and fraud[12]. In this paper ,here the most general definition, which is an outlying outlier, it appears to deviates to mark from other members of sample in which it occurs. To finding difficulties of outlier detection is the inherent difficulty in defining and quantifying the notion of outlier.

There are three type of learning algorithm, the mostly used unsupervised learning ,in unsupervised detection are suffers from high false alarm rate and low detection rate without labelled information. some are supervised learning[13] [14] it need large amount of training data which is difficult and taking long time and it is so difficult to find new type of outlier. And now a days semi-supervised learning is used to outlier detection. Semi-supervised outlier detection is continuously used unlabeled and labeled data , it can be improve the accuracy of outlier detection . The fuzzy C-means clustering method , FCM (Fuzzy C-means)[11] [15] it is one of most well known techniques in the cluster analysis by assign the fuzzy to each example and introduced a fuzziness weighted exponent. In this paper we are using the artificial neural network, inverse Gaussians algorithm as well as crisp value of the fuzzy set theory .Here we are considering the data set then divide these dataset into number group. After creating the data set we are going to find the mean of each cluster,

After applying the ANN (Artificial Neural Network) to each cluster the distinct point from the cluster are removed .Here we are using the numerical data of the experiments.

Related Work
IDS is split into two categories: misuse detection systems and anomaly detection systems. The misuse detection is used to identify intrusions that match known attack scenarios. However, anomaly detection is an attempt to search for malicious behavior that deviates from established normal patterns. Now our interesting is in anomaly detection.

In order to detect the intrusion, various approaches have been developed and proposed over the last decade [8]. In early stage, rule-based expert systems and statistical approaches are two typical ways to detect intrusion. A rule-based expert IDS can detect some well-known intrusions with high detection rate, but it is very difficult to detect new intrusions, and its signature database needs to be updated manually as well as frequently.

Statistical-based IDS, employs various statistical methods including principal component analysis, cluster and multivariate analysis, Bayesian analysis [3], and frequency and simple significance tests. But this type of IDS needs to collect enough data to build a complicated mathematical model, which is impractical in the case of complicated network traffic [11].

To solve the limitations of above methods, a number of data mining techniques have been introduced. Among these techniques, ANN is one of the most used techniques and has been successfully applied to intrusion detection According to different types of ANN, these techniques can be classified into the following three categories: supervised ANN-based intrusion detection, unsupervised ANN based intrusion detection, and hybrid ANN-based intrusion detection.

The current issue and full text archive of this journal is available at
http://aocrj.org/archive/
Academy of Contemporary Research Journal
V IV(II), 16-19, ISSN: 2305-865X

Overview of Fuzzy C-Means Clustering

fuzzy C-means clustering

This is generalization of HCM (Hard C-Means) which is one of the techniques in clustering by assign the fuzzy membership to each and every example and introducing fuzziness weighting exponent.

Problem statement

In fuzzy C-means clustering, there are number of points which are on the outlier of the cluster. The probabilities of these points are need to be find out. This is the main problem of the Fuzzy C-means clustering

Work System

Here we are using the data capture at real router in excel format dataset .this is the input of our project .The work is divided into three part.

A. *Fuzzy clustering technique*

The aim of the fuzzy cluster module is to partition of given set of data into clusters as well as it should have the following characteristics: 1) Homogeneity within the cluster 2) concerning data in same cluster and 3) Heterogeneity within the cluster where data belonging to different cluster should be as different as possible .Through this fuzzy clustering module ,the training set is clustered into several subset .Due to the fact that the complexity and size of every training subset is reduced. The effectiveness and efficiency of subsequent ANN module can be improved

The fuzzy cluster module is composed following step:

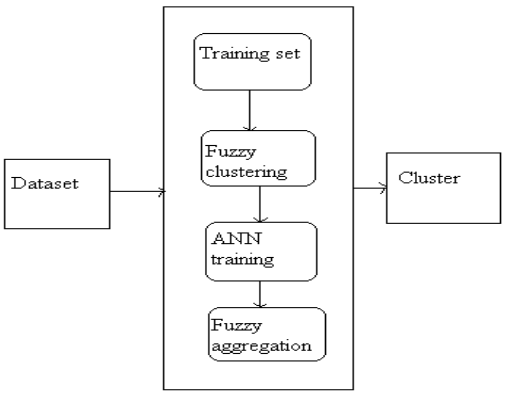
1) Initializes data set.
2) Calculating centers vectors.
3) Updating vectors.
4) Create the subset vector.

B. *ANN Training*

This module is going to learn the pattern of every subset .ANN is biological inspired from distributed computation .This is composed of simple processing units and connection between them. In this we, will go to classic feed-forward neural network trained with the back propagation algorithm to predict the intrusion. A feed-forward neural network have an input layer, an output layer with one or more hidden layers in between the output and input layers. The ANN function as follows: Each node i in the input layers has a signal xi as network's input, multiplied by the weight of value between the input layer and the hidden layer.

C. *Fuzzy aggregation*

The aim of this module is to aggregate different ANN's result and reduce the detection errors as every ANNi in ANN module only learns from the subset TRi because the error are nonlinear, in order to achieve the aims



.Fig 1 Data flow diagram

We use another ANN to understand the errors as follows

1) Let us take whole training set TR as data to input every trained ANNi and get the output
2) Form the input for new ANNi
3) Train the new ANN.

Set Theory

S = { N,A, Fc, At ,Ts , Fa }

Where

1. Let S = { } be as a System for Alert Identification and classification

2. Identify input as N
   Where N = Input Document of KDD cup data set and Live data captured at router side
   S= {N}

3. Identify A as Alert Identification and classification
   S = {N, A}

4. Identify process P

   S = {N, A, P}
   P = {Fc, At, Ts, Fa}

Where

Fc =Fuzzy Clustering
At =ANN Training
Ts =Training Set
Fa =Fuzzy Aggregation

Mathematical model

Fuzzy c Means Clustering Algorithm [FCM]

The FCM algorithm can be summarized by the following steps:

Step1: Initialize matrix $U=[U_{ij}]$ with the initial value U ;

Step2: At k-step: calculate the cluster prototype matrix

$$V^{(k)}=[v_i] \text{ with } U^{(k)};$$

Steps3:

Step4:

$$\text{if } \|U^{(k+1)}-U^{(k)}\|<\varepsilon \text{ then stop, or to step2.}$$

To sum up, the basic idea of the FCM algorithm is that use iterative method for solving equation (2) and (3), until a termination condition is met. Here, ε is the threshold of the termination condition.

Analysis and Result

The evaluation performance of FC-ANN approach, a series of experiments on Excel data where the data is captured on router side. The two different experiments on several real datasets are done and all experiments were performed in Window machine having configuration Intel (R) Pentium (R) 4, 1.73 GHz ,1GB RAM and windows 7/XP

The dataset contain about five million connection records as training data and about two million connections as test data. A dataset include a set of 41 features derived from each connection and a label will specifies the status of connection records as either a specific and normal attack types. This feature has all forms of discrete, continuous as well as symbolic variables.

With significantly varying range falling on four categories

1) This categories consists of the intrinsic feature of connection, which include the basic features of individual TCP connection. The time duration of connection, the types of protocol (UDP, TCP etc.) and the network services (http ,telnet etc.) are different feature.
2) This category content features within the connection suggested by domain knowledge are used to assess a payload of original TCP packets.
3) The same host features examine established connection in the last two seconds that has the same destination host as the current connection and calculate the statistics related protocol behavior, services etc.
4) Similar same service feature pay attention on the connection in the past two seconds that has the same service as the current connection

Like this, attacks are fall into four categories:

1) Denial of service (DOS): to make some computing or memory resources too busy to accept legitimate users access these resources.

2) Probe (PRB): host and port scans to collect information or find known vulnerabilities.

3) Remote to locate (R2L): Unauthorized access from a remote machine in order to exploit machine vulnerability

Update $U^{(k)}$, $U^{(k+1)}$;

4) User to root (U2R): unauthorized access to local super user privileged using system's susceptibility

However as the several instance for the U2R, PRB, and R2L attacks in the training set and test set is very low, these quantities is insufficient as a standard performance measure. Hence, if we are going to use these quantities as a measure for testing the performance of the systems, it could be biased. Only for these reasons, we give the precision, recall, and F-value which are not dependent on the size of the training and the testing samples. They are defined as follows:

Precision = TP/ (TP+FP)
Recall =TP/ (TP+FN)

Where, TP, FP, and FN are the number of true positives, false positives, and false negatives

To show the effectiveness of proposed system some experiments are conducted on java based windows machine. To measure the performance of the system we set the bench mark by selecting KDD CUP network data for efficient clustering and classification

To determine the performance of the system, we examined how many relevant clusters are formed based on the different network attacks using artificial neural network with fuzzy logic.

To measure this precision and recall are the best measuring techniques. So precision can be defined as the ratio of the number of relevant clusters retrieved to the total number of irrelevant as well as relevant clusters retrieved. It is usually expressed in the percentage.

Whereas Recall is the ratio of the total number of relevant clusters retrieved to the total number of relevant clusters in the database. It gives the information about the absolute accuracy of the system.

The advantage of having the two for measures like precision and recall is that one is more important than the other in many circumstances. In contrast, professional searchers and intelligence analysts are very concerned with trying to get as high recall as possible, and will tolerate low precision results in order to get it. Individually searching of their hard disks are often interested in high recall searches. Nevertheless, The two quantities clearly trade off against one another. Now we are going to calculate precision and recall.

Here we are assume:

• A = the number of relevant clusters retrieved,
• B = the number of relevant clusters not retrieved, and
• C = the number of irrelevant clusters retrieved.

So, Precision = (A / (A+ C)) *100
Recall = (A / (A+ B)) *100

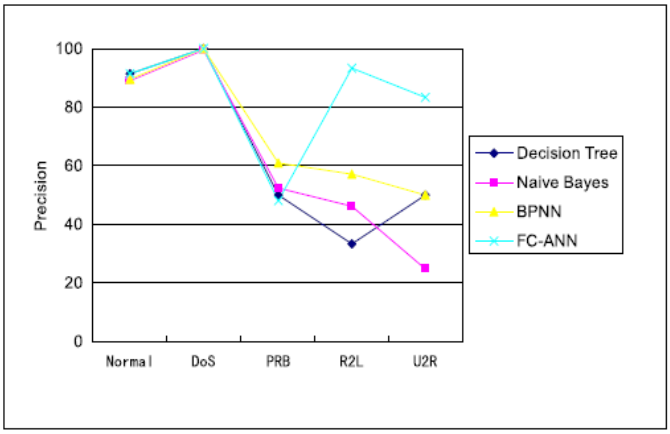


Fig. 2. The Retrieval average precision rate

In Fig. 2, here the tendency of average precision for the retrieved clusters are compared to the other systems.

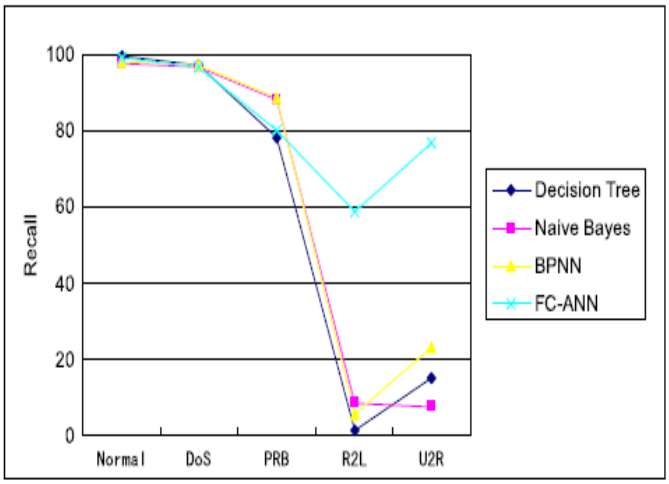
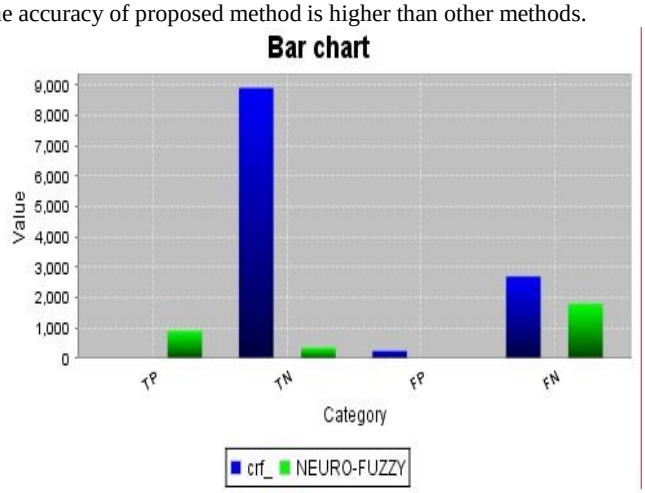


Fig.3. Retrieval average Recall rate

In Fig. 3, we observe that the tendency of average Recall for the retrieved clusters are high compared to other system. So this shows the accuracy of proposed method is higher than other methods.



Conclusion

This Prevention of security breaches fully using the existing security technologies is unrealistic. As a result, ID is an important component in network security. The IDS gives various advantages which helps in reducing manpower needed in monitoring, increasing detection efficiency, providing data that would otherwise not be available, helping the information security community learn about new vulnerabilities and providing legal evidence. In this paper, we propose a new intrusion detection approach, called Fuzzy C-means Artificial Neural Network , based on ANN and fuzzy clustering. Through this fuzzy clustering technique, the training set is divided to several homogenous subsets. Thus complexity of each and every sub training set is reduced and consequently the detection performance is increased. The experimental results are using the KDD CUP 1999 dataset demonstrates the effectiveness of our new approach especially for low-frequent attacks.

Future Scope

Here, we are going to use the ANN-Training so, in our system we use the random training dataset for calculation. In this system we remove the distinct point from the cluster,so that we can get the accurate cluster.

References

i. Anderson, J. P. (1980). Computer security threat monitoring and surveillance. Technical Report, Fort Washington, PA, USA.
ii. Anderson, J. (1995). An introduction to neural networks. Cambridge: MIT Press. Axelsson, S. (2003). The base-rate fallacy and the difficulty of intrusion detection. ACM Transaction on Information and System Security, 3, 186–205.
iii. Barbard, D., Wu, N., & Jajodia, S. (2001). Detecting novel network intrusions using Bayes estimators. In: Proceedings of the first SIAM international conference on data mining (pp. 1–17).
iv. Beghdad, R. (2008). Critical study of neural networks in detecting intrusions. Computers and Security, 27(5-6), 168–175.
v. Bezdek, J. C. (1973). Fuzzy mathematics in pattern classification. PhD thesis, Applied Math. Center, Cornell University Ithaca.
vi. Chen, Y. H., Abraham, A., & Yang, B. (2007). Hybrid flexible neural-tree-based intrusion detection systems. International Journal of Intelligent Systems, 22(4), 337–352.
vii. Chiu, S L. (1994). Fuzzy model identification based on cluster estimation. Journal of Intelligent and Fuzzy Systems, 2, 267–278.
viii. 8. Depren, O., Topallar, M., Anarim, E., & Ciliz, M. K. (2005). An intelligent intrusion detection system (IDS) for anomaly and misuse detection in CNs. Expert Systems with Applications, 29(4), 713–722.
ix. Dokas, P., Ertoz, L., Lazarevic, A., Srivastava, J., & Tan, P. N. (2002). Data mining for network intrusion detection. Proceeding of NGDM, 21–30.
x. 10.Endorf, C., Schultz, E., & Mellander, J. (2004). Intrusion detection and prevention.
xi. California: McGraw-Hill.
xii. Gordeev, M. (2000). Intrusion detection: Techniques and approaches. <http://
xiii. www.gosecure.ca/SecInfo/library/IDS/ids2.pdf> (accessed March 2009).
xiv. V. Barnett, T. Lewis, Outliers in Statistical Data, 3rd ed., John Wiley & Sons, New York, 1984.
xv. M. Markou, S. Singh, A neural network-based novelty detector for image sequence analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (10) (2006) 1664–1677.
xvi. S. Mukkamala, G.I. Janoski, A.H. Sung, Intrusion detection using neural networks and support vector machines, in: Proceedings of the 2002 International Joint Conference on Neural Networks, vol. 2, Honolulu, HI, USA, 2002, pp. 1702–1707.
xvii. J.C. Dunn, Some recent investigations of a new fuzzy partition algorithm and its application to pattern classification problems, Journal of Cybernetics 4 (1974) 1–15.