ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Intrusion Detection System Using Hybrid Approach by MLP and K-Means Clustering

Archana A. Kadam, Prof. S. P. Medhane

M.Tech Student, Bharati Vidyapeeth Deemed University, College of Engineering, Pune, Maharashtra, India

Assistant Professor, Bharati Vidyapeeth Deemed University, College of Engineering, Pune, Maharashtra, India

ABSTRACT: Intrusion Detection System (IDS) is a system that used for detecting the misuse of system i.e. malicious attacks. This system works as software for security management. Different techniques have been proposed by many researchers for Intrusion Detection System to achieve the better accuracy. In this paper we proposed hybrid approach to intrusion detection system. This hybrid approach make uses of k-means clustering and neural network Multi-Layer Perceptron (MLP) classification, this help to improve performance of the system. In this system we used KDD cup'99 dataset.

In this paper, we used clustering to divide the huge amount of data into different category. We have used k-means clustering for this purpose. In neural networks a set of nodes are used. These nodes are further used to label data for intrusion detection. In this way it gives solution to some of the neural networks inherent problems for classification to overcome like as the slow speed while labeling. Also the overhead of convergence and the burden of computation while classifying.

KEYWORDS: Intrusion detection, Hybrid Approach, Classification, Clustering, Neural Networking.

I. INTRODUCTION

Data mining techniques is generally used for IDS that implies investigation of the input dataset in an offline environment. Most popular data mining techniques are classification and clustering. The main reason behind use of classification techniques for IDS is estimation and prediction activity of IDS that is labeling of types of ID's. To discover similarities between data and for analysis clustering is used. Clustering is most valuable in intrusion detection system as malevolent activity should group together, unraveling itself from non-malicious activities. Following are some of the observations that motivated us to select k-means approach against its competent clustering approaches.

- The number of clusters desired is defined by user a priori and does not change.
- K-means works for unlabeled dataset consisting only numerical attributes.
- K-means algorithm is simple and it handles large data set very efficiently.

Neural Network (NN) is artificial, mathematical model inspired by biological neural networks. Artificial NN (ANN) for intrusion detection was first brought in as a substitute to statistical techniques in IDES intrusion detection expert system to model. ANNs have the capabilities of aligning the source data with its corresponding target output. This technique uses neural networks for intrusion detection for following reasons;

- The uniqueness such as high tolerance for noisy data, faster information processing, effective classification and the capacity of learning and self organization makes ANN flexible and powerful in IDS.
- The operation of IDS complies with intent of neural network models.
- The feature of dimensionality reduction and data visualization in neural networks can be helpful to reduce huge dimension of data records of a network connection.
- ANNs are skilled to handle little incomplete knowledge existing in IDS attacks.

II. RELATED WORK

For detecting the attacks, IDS is most commonly used. Intellectual IDS is equally capable as a dynamic defensive system which can adapt the changing traffic pattern dynamically. Many researchers have worked on Intrusion

Detection System to provide the finest output through their system. In a more simplified manner, the researchers have provided the IDS system including the artificial neural network and clustering techniques.

A more capable intrusion detection model is proposed by merging the skilled data mining techniques which consist of K-means clustering, Multilayer layer Perceptron (MLP) neural network and support vector machine (SVM), thereby improving the prediction of network intrusions. Finally, the SVM classifier is used for producing superior results for binary classification. Thus the best results are produced to prove the effectiveness of our model [4].

Here author proposed a technique in which first step is to divide the large amount of data into small subset of data. For this k-means clustering is used. Once the cluster is formed for classification or labeling dataset, of neural network set is used. This gives the idea about all the subspaces for intrusion detection separately. This subdivision of dataset is used for solving some neural networks inherent problem. These problems are slow in speed for labeling or converting and the burden of computation. Also, in this technique frequency called by the system is replaced by the call order of the system. System calls frequency represents the behavior of the program [5].

Two classification methods are presented in this paper i.e., radial basis function and multilayer Perceptron. Multilayer Perceptron group i.e layer with different nodes and radial basis function is used. Here we introduce hybrid architecture that includes a set of base classifiers for IDS. This leads to the enhanced performance of the system as compared to that of the use of existing classification methods. This technique may result into a significant improvement in the prediction accuracy of the intrusion detection. Here, for intrusion detection model we have used RBF and MLP in general. Performance comparison is also determined using intrusion detection datasets. We will get complementary feature of the base classifiers by using both RBF and MLP. Finally, we introduced hybrid architecture for intrusion detection model that involves group of base classifiers [6].

Here, we introduce Neural Network Committee Machine (NNCM), it consist Input Reduction System which is based on Intrusion Detection System and Principal Component Analysis (PCA) and these are represented by three level committee machine, each of this is based on Back-Propagation Neural Networks during working on this project we have used multiple methodologies, these are: The first one is PCA used for feature reduction, second is feed forward Multilayer Perceptron Neural Networks used for classification and the last one is Committee Machine for mechanism boosting purpose [7].

A new model named as Intrusion Detection Model is proposed which has three main parts: First is Input reduction system for reducing number of inputs from 41 to 13. Second is intrusion detection system and the last one is offline update system. SVM-based intrusion detection system combines a hierarchical clustering algorithm, a simple feature selection procedure, and the SVM technique. Unimportant features are eliminated from the training set through simple feature selection procedure so as to classify the network traffic data more accurately. SVM-based IDS uses BIRCH hierarchical clustering for data preprocessing. The BIRCH hierarchical clustering can provide highly optimized datasets, instead of original large dataset, to the SVM training [8].

This technique proposes a combination of the K-means clustering and Naïve Bayes classifiers (KM+NB), this means a hybrid learning approach. The evaluation and comparison of this approach was done by using KDD Cup'99 benchmark dataset. The Separation of the instances from the normal instance between potential Attacks is the best solution for this type of problem from a preliminary stage into different clusters. Subsequently, the clusters are further classified into more specific categories, namely Probe, R2L, U2R, DoS and Normal. The future scope of our system will be considering the extension of our project i.e. Signature based detection mechanism. This mechanism is better for detection of R2L and U2R attacks. [9].

Here, in our project, IDS merge with artificial neural network and k-means clustering for improving the system. To obtain a better result benchmark dataset was split into training and testing part and then cluster the dataset into five different divisions. After getting the cluster data it has been trained by the different Artificial Neural Networks functions as- Radial Basis Neural Networks (RBNN), Elman Neural Networks (ENN), Feed Forward Neural Networks (FFNN), Probabilistic Neural Networks (PNN) and Generalized Regression Neural Networks (GRNN). After implementing these, a best accuracy rate is chosen. Neural networks have been extensively used in anomaly detection system as well as misuse detection. Being a nonparametric model is its main advantage and it is easy to understand as compared to statistical methods. There are many artificial neural network functions used in our model namely Radial Basis Neural Networks, Elman Neural Networks, Feed Forward Neural Networks, Probabilistic Neural Networks and Generalized Regression Neural Networks [10].

III. PROPOSED ARCHITECTURE

The proposed system architecture has 3 stages. The system is described in detail. At very first the dataset is divided into two parts as train and test set with ratio of 60:40 respectively. Then with k-means clustering the train and test dataset cluster into attack and normal dataset. System trains for defect dataset clusters, and send it to the MLP on the last stage.

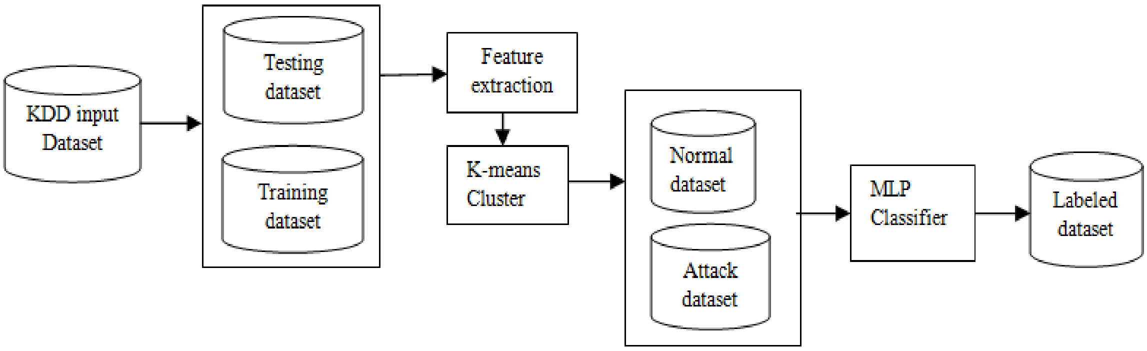

Fig1. Proposed System

A. System divided into three phases:

Phase 1: Divide the dataset into two sets- (a) Training dataset and (b) Testing dataset. Feature extraction using IG function for train dataset. Threshold value is set to extract 20 features from training dataset. K-mean clustering is used to training dataset to create cluster for normal and attack dataset.

Phase 2: Feature extraction using IG function for testing dataset. Same method used for training dataset. K-mean clustering is used to testing dataset to create cluster for normal and attack dataset.

Phase 3: Finally testing and training dataset given to MLP classifier to label the data stream. Result from the neural network functions is obtained as label of test dataset.

B. Division of Dataset

Here we used the KDD99 Dataset for experiment. Four main attack types are consider as- Probe, Dos, U2R (User to Root) and R2L (Remote to Local) and a Normal class. These names are initiated with the numeric value from 0-4. In this dataset the data stream consists of 41 features. Feature of the protocol type, the service type, and the flag is represented by the column 2, 3 and 4 respectively. Tcp, udp or icmp these are the some protocol type. The Service type has 66 different types of network services such as smtp and http. The flag has different 11 possible values like S2 or SE. Consider example of protocol type feature , i.e. 0 is assigned to tcp, 1 to udp, and 2 to the icmp symbol respectively.
[7].

C. Clustering the Dataset

As it is mentioned above in data mining to deal with large data clustering is the best process. There are many types of clustering techniques are available among that we used k-mean clustering. As it is the most effortlessly clustering approach for calculation [9,11].

K-means Clustering

K-means clustering is unsupervised data mining technique. It is one of the simplest methods. Also it is simplest to be implemented. It is fast, robust and understandable [6,1]. As the data set is distinct or wants to be well separated from each other the K-means clustering gives the best result in such case.

Algorithmic steps for k-means clustering [6]

Let A = {a1, a2, a3,……..,an} be the set of data points and B = {b1,b2,b3…….,bn} be the set of centers.
1) Selects cluster ‘C’ center randomly.
2) Calculate distance between cluster centers and each data point.
3) Assign the data points who are having minimum distance from cluster center as of all the cluster centers, to that cluster center.
4) By using the formula $vi=(1/ci)\ xicij=1$ recalculate the cluster center as new one.
Where ‘ci’ represents the number of data points in i number of cluster.
5) Recalculate the distance between newly obtains cluster centers each data point.
6) Stop, if no data point was reassigned otherwise repeat all steps from step 3.

D. Neural Network MLP Classification

A neural network is type of network as a set of distributed units following by a particular topology [3]. It creates connections between many types of processing elements as at different layer. But these layers are each parallel. It is extensively used in anomaly detection system as well as in misuse detection. It is easy to understand as compared to statistical methods. The main advantage of using this is it’s nonparametric model. Many artificial neural network functions are used in different model proposed by different author [6, 10]. Here we used MLP as the classifier algorithm. It works on basic principle of joining a every previous layer of neuron‘s output of the to the input of every neuron‘s to the next layer. One hidden layer must consist in MLP architecture. Due to the signal transition establishment through the network from input to output it is called as feed forward architecture.
Here in this project the neural networks classifier had used and it is applied testing just after the clustering phase to the training dataset and testing. Five specific categories have been done by this system as Dos, R2L, U2R, Normal and Probe. By using these five categories we can achieve detection and accuracy at very high range.

IV. RESULT AND DISCUSSION

This proposed system explained a comparison of SVM and MLP classification result with KDDCUP99 dataset and algorithm has been implemented in Weka.

A. Dataset & Selecting Features

In experiment, KDDCUP99 dataset is used to for evaluation. There are total 41 features in dataset, among that only 20 features are considered. And for features selection IG algorithm is used.
Result:

Data	Normal	DOS	Probe	R2L	U2R
Instance	2841.0	13122.0	600.0	205.0	103.0
Percentage	16.84 %	77.78 %	3.56 %	1.22 %	0.61 %
Accuracy	87.9 %	99.86 %	86.46 %	77.95 %	73.05 %

Table 1. Result of MLP Classifier

V. CONCLUSION AND FUTURE WORK

By using this hybrid approach we are able to improve the performance of system. So for improving the accuracy better feature selection techniques are that IG is applied. We have used the clustering technique for identifying the low frequent data for more specification.
It could be more accurate and effective in future if we work more on the following future works:

1. This research elaborate work for attacks we have known only. In future it will be more encouraging if we work for the novel attacks by enhancing the present system.

2. Clustering has different types of methods and that can be applied to improve our work. In future it may be excellent that for assessment if we calculate the accuracy for different cluster.

REFERENCES

1. Jonathon Ng, Deepti Joshi, Shankar M. Banik “Applying Data Mining Techniques to Intrusion Detection” 12th International Conference on Information Technology, 2015 IEEE Computer Society.

2. Nutakarn Mongkonchai, Phet Aimtongkham, Kasidit Wijitsopon and Kanokmon Rujirakul, “An Evaluation of Data Mining Classification Models for Network Intrusion Detection” Applied Network Technology (ANT) Laboratory Department of Computer Science, 2014 IEEE.

3. Mazyar Mohammadi Lisehroodi, Zaiton Muda, and Warusia Yassin, “A HYBRID FRAMEWORK BASED ON NEURAL NETWORK MLP AND K-MEANS CLUSTERING FOR INTRUSION DETECTION SYSTEM” 4th International Conference on Computing and Informatics, August 2013.

4. A.M. Chandrashekhar and K. Raghuveer, “Amalgamation of K-means Clustering Algorithm with Standard MLP and SVM Based Neural Networks to Implement Network Intrusion Detection System”, Advanced Computing, Networking and Informatics, Volume 2, Springer International Publishing Switzerland 2014.

5. Hongying Zheng, Lin Ni, Di Xiao, “Intrusion Detection Based on MLP Neural Networks and K-Means Algorithm”, Second International Symposium on Neural Networks, Volume 3498, Springer-Verlag Berlin Heidelberg, June 2005,

6. M. Govindarajan , RM. Chandrasekaran, “Intrusion detection using neural based hybrid classification methods”, Computer Networks, Volume 55, Issue 8, 1 June 2011, Pages 1662–1671, at ScienceDirect®.

7. Alma Husagic-Selman, Rasit Koker, Suvad Selman, “Intrusion Detection using Neural Network Committee Machine”, XXIV International Conference on Information, Communication and Automation Technologies (ICAT), 2013 IEEE.

8. Shi-Jinn Horng, Ming-Yang Su, Yuan-Hsin Chen, Tzong-Wann Kao, Rong Jian Chen, Jui-Lin Lai, Citra Dwi Perkasa, “A novel intrusion detection system based on hierarchical clustering and support vector machines”, Expert Systems with Applications, Volume 38, Issue 1, January 2011, Pages 306–313 at ScienceDirect®.

9. Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir, “K-Means Clustering and Naive Bayes Classification for Intrusion Detection”, Journal of IT in Asia, Volume 4, may 2014.

10. Wasima Matin Tammi, Noor Ahmed Biswas, Ziad Nasim, Faisal Muhammad Shah, “Artificial Neural Network based System for Intrusion Detection using Clustering on Different Feature Selection”, International Journal of Computer Applications, Volume 126, No.12, Pages 0975 – 8887, September 2015

11. K. M. Faraoun and A. Boukelif, “Neural Networks Learning Improvement using the K-Means Clustering Algorithm to Detect Network Intrusions”, International Journal of Computational Intelligence, Volume 3, Number 2005.