International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 5 Issue 9|| September 2016 || PP. 21-*24*

Intrusion Detection System for Classification of Attacks with Cross Validation

Akhilesh Kumar Shrivas[1], Prabhat Kumar Mishra[2]
[1, 2]*(Department of IT, Dr. C. V. Raman University, Kota, Bilaspur (C.G.), India)*

Abstract: *Now days, due to rapidly uses of internet, the patterns of network attacks are increasing. There are various organizations and institutes are using internet and access or share the sensitive information in network. To protect information from unauthorized or intruders is one of the important issues. In this paper, we have used decision tree techniques like C4.5 and CART as classifier for classification of attacks. We have proposed an ensemble model that is combination of C4.5 and Classification and Regression Tree (CART) as robust classifier for classification of attacks. We have used NSL-KDD data set with binary and multiclass problem with 10-fold cross validation. The proposed ensemble model gives satisfactory accuracy as 99.67% and 99.53% in case of binary class and multiclass NSL-KDD data set respectively.*

Keywords: *Intrusion Detection System(IDS), Cross Validation, Decision Tree, Classification.*

I. Introduction

As Rapid development of internet and intranet, protecting of information or data from intruder is one of the important issues. Basically IDS is a classifier that is used to identify the set of malicious or suspicious action and classify the suspicious events from normal data. IDS[1] is a powerful security tools in computer system environments and deploy in computer network to protect attacks by intruders or unauthorized person. Intrusion detection system [2] can be categorized into two types: Misuse intrusion detection and Anomaly intrusion detection. Misuse intrusion detection technique is based on supervised learning that means training the pattern of attacks by supervised learning of labeled data while Anomaly based intrusion detection system is based on unsupervised learning that means this technique establish normal usages patterns. Misuse intrusion detection cannot identify new types of attacks that have never occur in training data while Anomaly based intrusion detection system can detect the unseen intrusions by investigating their deviation from the normal patterns. There are various authors have worked in the field of IDS. S. Lakhina et al. [8] have proposed new hybrid technique PCANNA (Principal Component Analysis Neural Network Algorithm) as intrusion detection system for classification of attacks. They have used Principle Component Analysis (PCA) as feature selection technique on NSL-KDD data set and reduced data set is feed to neural network. The proposed model given better classification accuracy with 8 numbers of features of NSL-KDD data set which has reduced training time by 40% and testing time by 78.5%. V. Balon Canedo, et al. [9] proposed a new KDD winner method consisting of discretizations, filters and various classifiers like Naive Bayes (NB) and C4.5 to develop robust IDS. The proposed classifier given high accuracy i.e. 99.45% compare to others. H. Altwaijry et al. [10] have suggested bayesian network to improve the accuracy of R2L type of attack. Experiment done with different feature subset of KDD99 data set and given better results for R2L type of attack with detection rate 85.35%. A. K. Shrivas et al. [11] haves suggested ensemble of Artificial neural network (ANN) and Bayes Net for classification of attacks and Normal data in case of NSL-KDD data set. The proposed technique given 98.07% with 35 features in case of Gain ratio feature selection technique. Y. B. Bhavsar et al. [12] have proposed Support Vector Machine(SVM) with different kernel function for classification of various types of attacks. The proposed SVM with RBF kernel function given better classification accuracy as 98.57% with 10-fold cross validation in NSL-KDD data set. L. Dhanabal et al. [13] have used NSL-KDD data set and applied on J48, Support Vector Machine (SVM) and Naïve Bayes for classification of attacks and normal samples. C4.5 has given best accuracy in case of all types of attacks and normal data with 6 feature subset.

II. Proposed Model

The Fig.1 shows that the proposed architecture of development of robust IDS. This figure shows that NSL-KDD data set divides into 10 fold or part means $1/10^{th}$ parts is used as testing data and rest of the part is used as training samples. This process continues 10^{th} times until each partition uses as testing samples and rest of partitions use as training samples. In this model, firstly 1 to 9^{th} partitions are used as trained the ensemble of C4.5 and CART model and 10^{th} partition is used as testing the model. Secondly, other one part of data set is used as testing and rest of parts are used as training the ensemble of C4.5 and CART model. Similarly other part is used as testing and rest of parts are used as training. Finally, overall testing accuracy is calculated which is robust compare to accuracy of individual models i.e.C4.5 and CART.

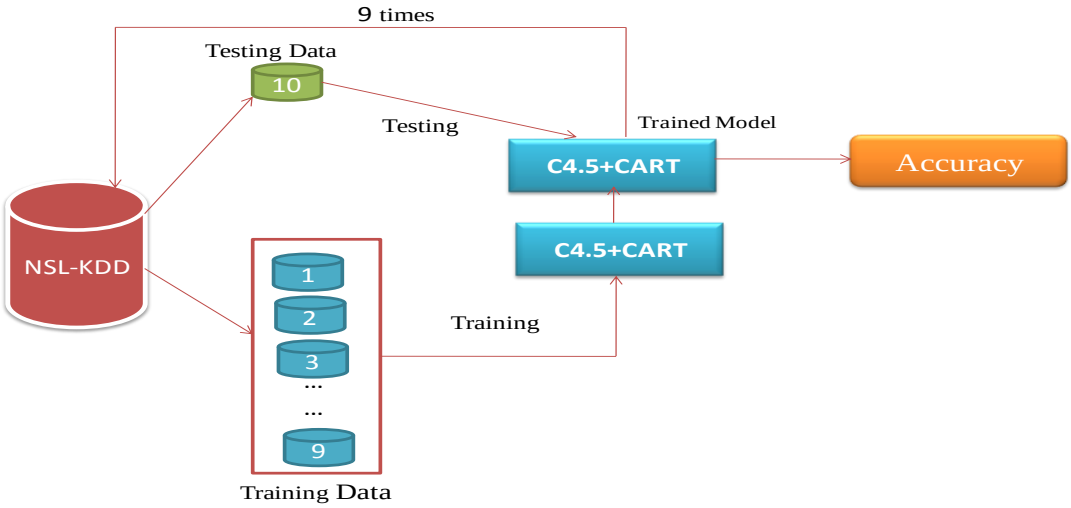

Fig. 1: Proposed Architecture

The Pseudo code of proposed model is shown below:

Input

F_n= NSL-KDD data set with 41 features f_1,f_2,f_3…..f_{41}

Output

Develop robust model M

Proposed Model

1. Start
2. Supply F_n to C4.5 and CART with 10-fold cross validation.
3. A_1= Accuracy of C4.5 model.
4. A_2= Accuracy of CART model.
5. E= Ensemble of C4.5 and CART with 10-fold cross validation.
6. A_3= Accuracy of E.
7. Compare the accuracy of A_1, A_2, and A_3.
8. Select the best Model M= A_3.
9. End

III. Data Set

NSL- KDD [7] is One of the publicly available data set for the evaluation of intrusion detection system which is solving some of the inherent problems of the KDD'99 data set. One of the most important efficiencies in the KDD data set is the huge number of redundant records, which causes the learning algorithms to be biased towards the frequent records, and thus prevent them from learning infrequent records which are usually more harmful to networks such as U2R and R2L attacks. This research work have used 25192 records of NSL-KDD data set .In this research work, we have used two types of NSL-KDD data set : one is binary class and second is multiclass data set. Binary class contains normal and attack types of samples while multiclass data set contains one type of normal and four type of attacks data Like DoS, R2L, U2R and Probe. All the features of NSL-KDD data set same as features of KDD99 data set. Table 1 shows that sample size of two class problem and table 2 shows that sample size of multiclass problem.

Table 1: Sample size of two class problem

Category of class	**Number of instances**
Normal	13449
Attacks	11743
Total	**25192**

Table 2: Sample size for multiclass problem

Category of class	**Number of instances**
Normal	13449
DoS	9234
R2L	209
U2R	11
Probe	2289
Total	**25192**

IV. Decision Tree Technique

Decision tree [3] is the most popular data mining technique. The most common data mining task for a decision tree is classification. The basic idea of a decision tree is to split our data recursively into subsets so that each subset contains more or less homogeneous states of our target variable (predictable attribute). At each split in the tree, all input attributes are evaluated for their impact on the predictable attribute. When this recursive process is completed, a decision tre is formed. This research work has used CART and C4.5 decision tree for classification of various types of attacks.

CART [4] is one of the popular methods of building decision tree in the machine learning community. CART builds a binary decision tree by splitting the record at each node, according to a function of a single attribute. It uses the gini index for determining the best split. The initial split produces the nodes, each of which we now attempt to split in the same manner as the root node. Once again, we examine the entire input field to find the candidate splitters. If no split can be found then significantly decreases the diversity of a given node, we label it as a leaf node. Eventually, only leaf nodes remain and we have grown the full decision tree. The full tree may generally not be the tree that does the best job of classifying a new set of records, because of overfitting.

C4.5 [4] is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees and rule derivation. In building a decision tree, we can deal with training sets that have records with unknown attributes values by evaluating the gain, or the gain ratio, for an attribute values are available. We can classify the records that have unknown attribute value by estimating the probability of the various possible results. Unlike, CART, which generates a binary decision tree, C4.5 produces tree with variable branches per node. When a discrete variable is chosen as the splitting attribute in C4.5, there will be one branch for each value of the attribute.

V. Ensemble Model

An ensemble model [5] combines the output of several classifier produced by weak learner into a single composite classification. It can be used to reduce the error of any weak learning algorithm. The purpose of combining all these classifier together is to build a hybrid model which will improve classification accuracy as compared to each individual classifier. In this research work, we have used voting scheme to ensemble the two models. This research work have used ensemble the C4.5 and CART technique to develop the robust model.

VI. Experimental Work

The experiment is carried out using NetBeans IDE 8.0.2 and coding is done in java with WEKA data mining tool [6] library. The main objective of this research work is to develop a robust classifier for classification of various types of attacks. We have used NSL-KDD data set that is collected from UCI repository. We have categorized NSL-KDD data set into two sections: NSL-KDD with binary class and NSL-KDD with multiclass problem. We have used 10-fold cross validation technique to partition the data set. These two categories of data set have applied into decision tree techniques like C4.5 and CART to develop a classifier for classification of attacks ,but individuals model are not capable to identify the false alarm rate with better accuracy. To overcome the problem of individual models, we have ensemble C4.5 and CART technique to increase the accuracy of model. Table 3 shows that accuracy of individuals and ensemble model in case of binary and multi class data with 10-fold cross validation. Individual models C4.5 give 99.56% and 99.46% accuracy in case of binary and multiclass problem respectively while CART gives 99.66% and 99.46% accuracy in case of binary and multiclass problem respectively. Our proposed model as ensemble of C4.5 and CART give 99.67% and 99.53% accuracy in case of binary and multiclass problem respectively. Fig 2 shows that accuracy of individuals and ensemble model with 10-fold cross validation. Finally our proposed ensemble model is efficient model for classification of attacks and normal data.

Table 3: Accuracy of Model with 10-fold cross validation of NSL-KDD data set

Model	Binary Class	Multiclass
C4.5	99.56%	99.46%
CART	99.66%	99.46%
C4.5+CART	99.67%	99.53%

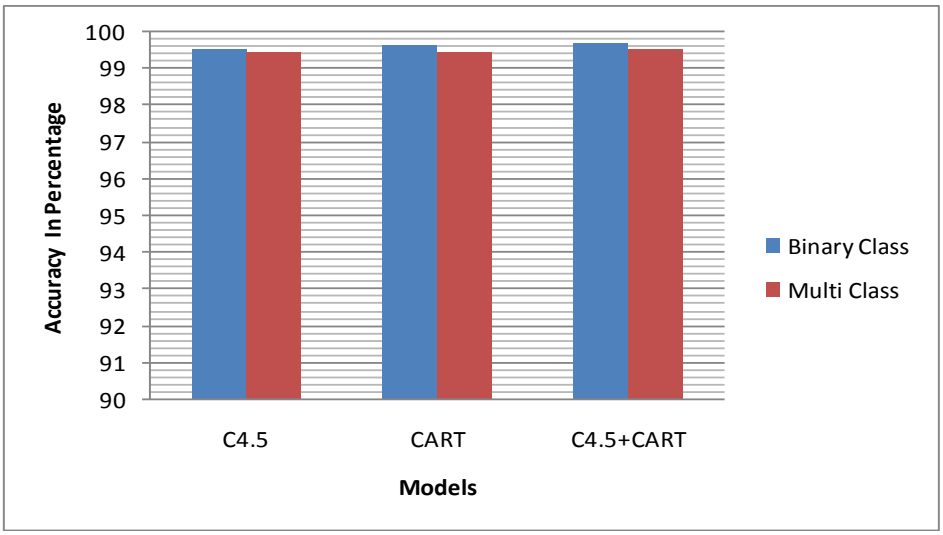

Fig. 2: Accuracy of various models with 10-fold cross validation

VII. Conclusion And Future Work

Intrusion detection techniques are important to protect the system form intruders or malicious behaviors. In this paper, developed a new ensemble model that is combination of C4.5 and CART for classification of normal and different types of attacks. The main advantage of new developed model is that it achieved high classification accuracy compare to individuals model as C4.5 and CART. The proposed ensemble model identifies the false alarm rate with high accuracy.

In future, we can apply the various raking based and our own developed feature selection techniques to computationally increase the performance of model. Feature selection is remove the irrelevant data form original data set and select the important data which is used to develop the efficient model. We will also use other optimization techniques like Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) techniques to develop computationally efficient and robust model.

References

[1]. N. Y. Jan,, S. C., S. S. Tseng and N. P. Lin, A decision support system for constructing an alert classification model, Expert Systems with Applications, 36 ,2009, 11145–11155.

[2]. J. Z. Lei , Ghorbani and , A. Ali , Improved Competitive Learning Neural Networks for Network Intrusion and Fraud Detection, Neurocomputing, 75, 2012, 135-145.

[3]. Z. Tang and J. Maclennan, Data Mining with SQL Server 2005.Willey Publishing, Inc, USA, ISBN: 13: 978-0-471-46261-3, 2005.

[4]. A. K. Pujari, Data Mining Techniques, Universities Press (India) Private Limited, 4th ed., ISBN: 81-7371-380-4, 2001.

[5]. M. Pal , Ensemble Learning with Decision Tree for Remote Sensing Classification, World Academy of Science, Engineering and Technology. 36, 2007, 258-260.

[6]. WEKA Data Mining Tools: http://www.cs.waikato,ac.nz/~ml/weka/ (Browsing date: June 2016).

[7]. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets.html].

[8]. S. Lakhina, S. Joseph and B. Verma , Feature Reduction using Principal Component Analysis for Effective Anomaly Based Intrusion Detection on NSL-KDD, International Journal of Engineering Science and Technology, 2(6), 2010, 1790-1799.

[9]. V. Bolon Canedo, N. Sanchez Marono, and , A. Alonso Betanzos, Feature Selection and Classification in Multiple Class Datasets: An Application to KDDCup 99 Dataset, Expert Systems with Applications, 38,2011, 5947-5957.

[10]. H. Altwaijry, and S. Algarny, Bayesian based Intrusion Detection System, Journal of King Saud University Computer and Information Sciences. 24,2012, 1-6.

[11]. A. K. Shrivas and A. K. Dewangan, An Ensemble Model for Classification of Attacks with Feature Selection based on KDD99 and NSL-KDD Data Set, International Journal of Computer Applications,99,2014,8-13.

[12]. Y. B. Bhavsar and K. C.Waghmare, Intrusion Detection System Using Data Mining Technique: Support Vector Machine, International Journal of Emerging Technology and Advanced Engineering, 3, 2013, 581-586.

[13]. L. Dhanabal and Dr. S.P. Shantharajah, A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms, International Journal of Advanced Research in Computer and Communication Engineering, 4, 2015, 446-452.