Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 49 (2015) 92 – 98

Procedia Computer Science

ICAC3'15

Intrusion Detection System Using Bagging with Partial Decision TreeBase Classifier

D.P.Gaikwad[a*],Ravindra C.Thool[b]

[a]*Department of Computer Enginnering,AISSMS College of Engineering, Pune-410014,India*

[b]***Department of Information Technology, SGGS Instutute of Engineering & Technology,Nanded-431606,India***

Abstract

Intrusion Detection System has become an essential part of the computer network security. It is used to detect, identify and track the intruders in the computer network. Intrusion Detection Technology which provides highest classification accuracy and lowest false positive is required. Many researchers are involved to find out and propose Intrusion detection technology which provides the better classification accuracy and less training time. The traditional Intrusion Detection system exhibits low detection accuracy and high false alarm rate. Now a day,an Ensemble method of machine learning is widely used to implement intrusion detection system. By analyzing Ensemble method of machine learning and intrusion detection system in this paper, we make use of Bagging Ensemble method to implement Intrusion Detection system. The Partial Decision Tree is used as a base classifier due to its simplicity.The selections of relevant features are required to improve the accuracy of the classifier. The relevant features are selected based on their vitality for each type of attacks. The dimension of input feature space is reduced from 41 to 15 features using Genetic Algorithm.The proposed intrusion detection system is evaluated in terms of classification accuracy, true positives, false positive and model building time. It was observed that proposed system achieved the highest classification accuracy of 99.7166 % using cross validation. It exhibits higher classification accuracy than all classifiers except C4.5 classifier on test dataset.The Intrusion Detection system is simple and accurate due to simplicity of Partial Decision Tree.

Peer-review under responsibility of organizing committee of the 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15)

Keywords: Ensemble;Machine Learning;Model Building Time;Partial Decision Tree;Genetic Algorithm

*D.P.Gaikwad Tel: +91-09822609276
E-mail address: dp.g@rediffmail.com

Peer-review under responsibility of organizing committee of the 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15)
doi:10.1016/j.procs.2015.04.231

1. Introduction

Internet is widely developing in world as communication media. Due to rapidly development of Internet, more and more systemsencounter the intruders in network. The intruder can access, manipulate and disable computer system through Internet. Therefore, the securities of computer system in network have become an essential requirement in computer network system.The intrusion detection system is used to detect unauthorized access to computer system in network. It is used to prevent such unwanted access, manipulation of computer functionality and external penetration in the private organization. It is used to detect, identify and prevent all types of network attacks in network environment. The malicious activities are detected by analyzing the packets to prevent damage from attack. Generally, intrusion detection techniques are categorized into two methods: misuse and anomaly detection. Misuse detection method is also called as signature based detection system. It is used to detect attacks based on the known pattern of attack. They are used to detect known attacks effectively with low errors. They are unable to detect unknown attacks because new attack do not have similar pattern to known attack. Anomaly detection technique is profile based which analyses normal traffic. It detects unknown packets in network effectively, but they are not so effective in detection rate. They also provide high false positive rates. To resolve the disadvantages of the anomaly detection technique of intrusion detection, machine learning technique have also been used by many researcher. The detection performance of the machine learning depends on the technique of machine learning. The ensemble method of machine learning is more efficient which can reduce the false alarms and increase the classification accuracy. There are three methods of ensemble methods: Bagging, Boosting and Stacking [1] [2]. Bagging and boosting ensemble methods are widely used to implement the intrusion detection system as compared to Stacking. The stacking of weak classifier require more time, so they are not practically effective for intrusion detection. In this paper, we make use of bagging method with Partial Decision Tree as weak classifier to implement intrusion detection system. The selection of relevant features from dataset is required to improve the classification accuracy and reduce the false positives. The relevant features are selected based on their vitality to identify the types of attacks. The vitalities of features are determined based on our literature survey and experience. Table 1 shows the list of relevant features used to train and test the proposed intrusion detection system. The performance of the system is evaluated in term of false positive, classification accuracy and model building time. The rest of this paper is organized as follows. Section II surveys some previous work on Machine Learning based intrusion detection system. Section III introduces Partial Decision Tree base classifiers. In Section IV, Overview of Proposed Intrusion Detection System is given. Section V is dedicated to experimental results and discussions. Finally, Section VI concludes the paper.

2. Related Works

Shrinivasu and P.S.Avadhani [3] have proposed GA-NN based intrusion detection system. Genetic Algorithm Weight Extraction Algorithm is used to extract and optimize the weights between the neurons of ANN to identify the intrusions effectively. Li Hanguang, Ni Yu [4] have used Apriori algorithm which generate to identify is a variety of attacks, improves the overall performance of the detection system. Gisung Kim, Seungmin Lee and Sehun Kim [5] have proposed hybrid intrusion detection method. The method integrates the anomaly and misuse detection in hierarchical manner. A misuse detection model is based on c4.5 classifier. The one-class SVM models are trained using subsets of training dataset which reduces false positives effectively.Wei Wang et al., [6] have proposed automatic intrusion detection system using dynamic clustering method. It is online and adaptive intrusion detection system. Wenying Feng et al., [7] have combined SVM method and Clustering based on Self-Organized Ant Colony Network to implement the intrusion detection system. The proposed method takes the advantages of both SVM and Clustering based on Self-Organized Ant Colony Network which avoids their weaknesses.Fangjun Kuanga et al., [8] have proposed a Novel hybrid KPCA SVM with GAs model for intrusion detection. In this model, KPCA is used to extract the principal features of intrusion detection data. The SVM multi-layer classifier is used to identify an attack.

3. Features Selection and Preliminaries of Partial Decision Tree

In this section, we provide a brief explanation of feature selection method used in this paper. The preliminary of Partial Decision Tree is discussed in brief as follows:

3.1. Feature Selection

The online available datasets provided by DARPA 1998, NSL-KDD99 and KDD99 are mostly used as training dataset in intrusion detection system. In this paper, the NSL-KDD99 dataset is used to carry out experiments. The NSL-KDD dataset have suggested 41 features. If we use all features in dataset for training, then it take more time for model building and they also can affect the accuracy. To avoid this, in pre-processing step of intrusion detection features selection is require reducing dimension, boosting generalization capability, accelerating learning and enhances model interpretation [9].In this paper, Genetic Algorithm is applied on NSL_KDD dataset to select relevant feature. The Genetic Algorithm selects 15 features out of features from dataset. In table 1, all selected features are listed.

Table 1. List of Features

Sr.No.	*Features*
1	Flag
2	Src_bytes
3	Dst_bytes
4	Wrong_fragment:
5	Hot
6	Logged_in
7	Num_file_creations
8	Count
9	Srv_serror_rate
10	Same_srv_rate
11	Diff_srv_rate
12	Dst_host_count
13	Dst_host_srv_diff_host_rate
14	Dst_host_serror_rate
15	Dst_host_srv_serror_rate

3.2. Preliminaries of Partial Decision Tree rule Learner

There are many schemes to generate rules from decision trees. The C4.5 and RIPPER are two main schemes for rule learning. The both scheme operate in two stages. The c4.5 first induces an initial rule set and then it refine rule set using complex optimization stage by discarding the individual rule. The RIPPER do same thing by adjusting individual rules. These two schemes can be combined to produce good rule sets. This combination of two scheme of rule learning is called as Partial Decision Tree (PART). This combined scheme does not require any complex optimization stage. The algorithm to combine C4.5 and RIPPER is very simple, effective and straightforward. Initially, it built a pruned decision tree for current set of instances. The leaf (best) with largest coverage is converted into rule, and decision tree is discarded by removing covered instances form training dataset. This process is repeated for all set of instances of training dataset. This process is called as separate-and–conquers strategy. PART algorithm produces rule sets which are more accurate than RIPPER's rule set. PART's rule sets are as accurate as C4.5's rule set and the size of rule sets of PART are of same size of C4.5 rule set. The performance of PART is fast because it does not need any post processing[10].

4. Overview of Proposed Intrusion Detection System

4.1. Architecture of IDS

The system architecture of proposed intrusion detection system is given in Fig.1. In this section we present two contributions; one is to select the relevant features from NSL_KDD99 dataset and second is to reduce the false positives. The Genetic Algorithm is used to select the relevant features. In second contribution, the bagging ensemble method machine learning is used to reduce the variance. The bagging method with Partial Decision tree as a base classifier is used to reduce the false positive and increase the classification accuracy. Once training is completed, the model with rule set is built. The performance of rule model is evaluated using crossvalidation of 10-fold and test dataset.

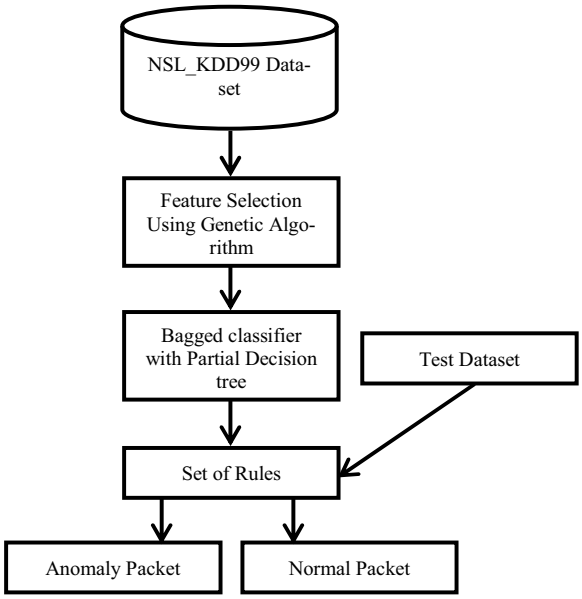

Fig.1. System Architecture of Intrusion Detection System.

4.2. Algorithm of Proposed IDS

The bagging is a kind of voting algorithm which takes a base classifier and training set as input. It runs multiple times by changing the distribution of instances in training dataset. Each trained base classifier then combined to generate classifier that is used to classify the test dataset. Bagging is also called as Bootstrap Aggregating. In this voting method, classifiers are generated by different bootstrap samples S_m. The samples are generated by uniform sampling n instances from training set with replacement. The classifiers C_1;C_2 ;C_3;:::: C_m are built using M bootstrap samples S_1;S_2 ;S_3;:::: S_m. The final classifier C* is built from the C_1; C2; C3;C_m whose output the most often predicted by the base classifier. The basic procedure for proposed intrusion detection system is summarized in Algorithm 1. The main reason for choosing PART is that it is simple, effective and straightforward decision tree.

5. Experimental Results and Discussion

In this section, the performance of Bagging with Partial Decision Tree as a base classifier and Genetic algorithm are presented. All experiments are performed by using an Intel(R) CORE™ i5-3210M CPU @ 2.50GHz, Installed 8GB RAM and 32 bit Operating system. Feature selection using Genetic Algorithm plays an important role in building classification systems. It reduces the dimension of data, model building time and lowers the computation costs.

Fifteen features are selected before passing the data sets to the Bagging of classifiers. The performance results of classifier using cross validation are listed in Table 2.

Algorithm 1: Bagging of PART for IDS

Input: NSL_KDD dataset, 15 features
Begin:
1. Let m=number of bootstrap samples
2. for i =1 to m do
3. Create a bootstrap samples S_1;S_2 ;S_3;:::: S_m(Sample with Replacement)
4. Train Partial Decision Tree as a base classifier (C_i) on bootstrap samples S_m
5. end for
6. C^* (x)= arrgmax Σ_i $\delta(C_i (x)=y)$ (the most often predicted label y)

End.
Output: Trained C* classifier

According to Table 2 and Fig 2, bagging with Partial Decision Tree exhibit highest accuracy of 99.7166 % on cross validation.In Fig 3, RMSE, True Positive and False positive rates are given. The performance results of classifier on test dataset are listed in table 3. According to Table 3 and Fig 4, bagging with Partial Decision Tree exhibit classification accuracy of 78.3712% on test dataset. The classification accuracy of C4.5 tree is more than Bagged PART.In Fig 5, RMSE, True Positive and False positive rates are given on test dataset.

Table 2. Performance Analysis of classifiers using Cross Validation.

Classifiers	RMSE	True Positive Rate	False Positives	Model Building Time Sec.	Accuracy In %
Naïve Bays	0.3148	0.896	0.114	42.74	89.6002
PART	0.054	0.997	0.003	278.96	99.6634
C4.5/J48	0.0517	0.997	0.003	175.96	99.6991
Bagging(Naïve Bays)	0.3112	0.895	0.114	225.25	89.4882
Bagging(PART)	0.0477	0.997	0.003	1342.42	99.7166
Bagging(C4.5)	0.0472	0.997	0.003	1686.8	99.7158

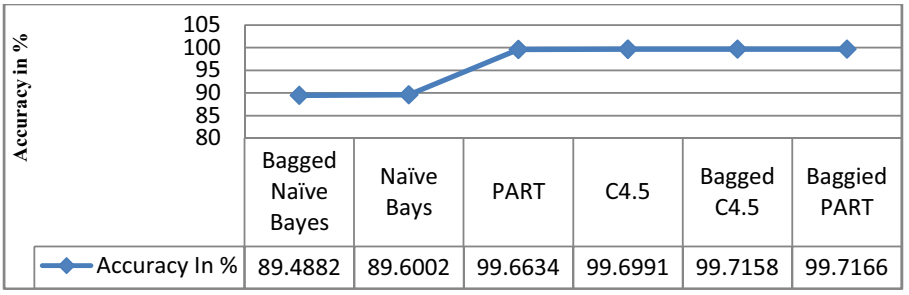

Fig.2 Classification Accuracy of Classifiers using Cross Validation.

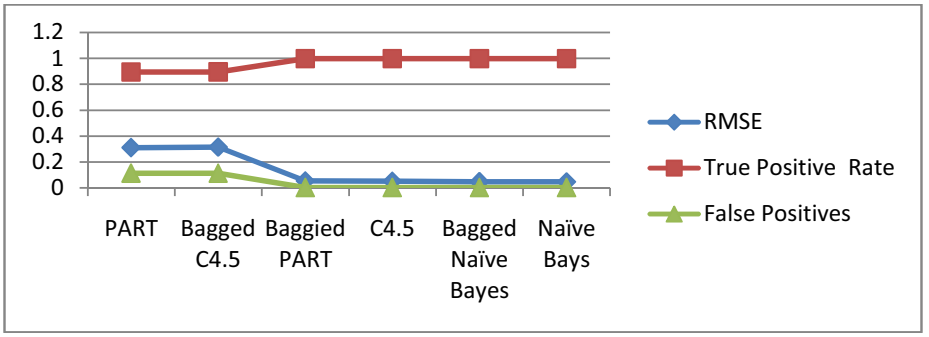

Fig.3.True Positives and False Positives with RMSE of Classifiers using Cross Validation.

Table 3. Performance Analysis of Classifiers on Test Dataset.

Classifier	*RMSE*	*True Positive Rate*	*False Positives*	*Model Build-ing Time Sec.*	*Accuracy In %*
Naïve Bays	0.5072	0.74	0.212	42.92	73.9798
PART	0.4664	0.778	0.176	274	77.7901
C4.5	0.4534	0.791	0.165	176.05	79.0809
Bagging(Naïve Bays)	0.5068	0.74	0.212	220.62	73.9798
Bagging(PART)	0.4418	0.784	0.172	1589.86	78.3712
Bagging(C4.5)	0.4552	0.779	0.174	1795.94	77.8699

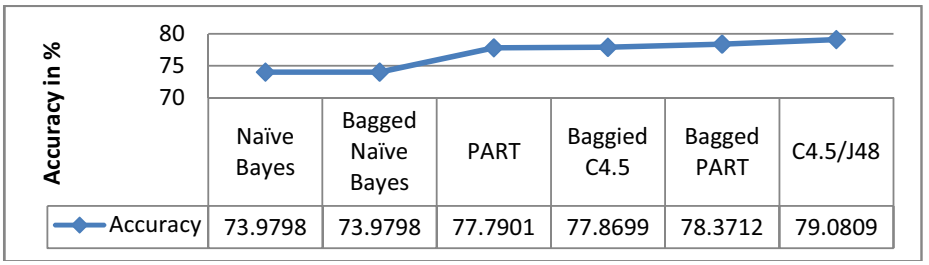

Fig.4. Classification Accuracy of Classifiers on Test Dataset.

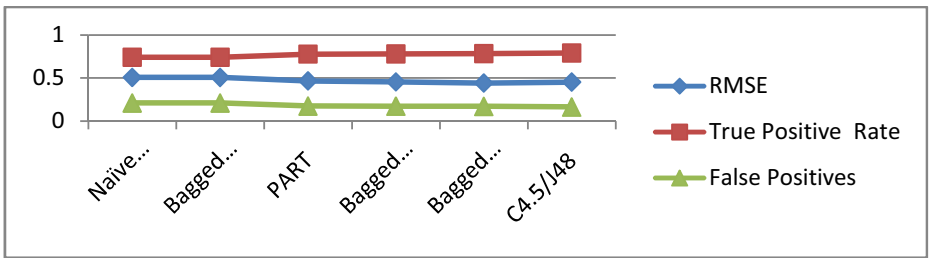

Fig.5.True Positives and False Positives with RMSE of Classifiers on Test Dataset.

6. Conclusions

In this paper, we discussed current approaches of intrusion detection system using machine learning techniques. Now a day, the Ensemble method of machine learning is widely used in pattern classification. In Ensemble method, a base classifier repeatedly trained on subset of training dataset. Any decision tree can be used as a base classifier in ensemble method. In this paper, the bagging method of machine learning is used to implement the intrusion detection system. We make use of Partial Decision Tree rule learner as a base classifier due its simplicity. Genetic algorithm is used to select relevant features from NSL_KDD99 dataset which have reduced the model building time and have improvedthe performance of proposed intrusion detection system. The performance of proposed IDS is evaluated in terms of RMSE, True Positive, False Positive rates and classification accuracy. The experimental resultsshow that bagging with Partial Decision Tree exhibit highest classification accuracy of 99.7166 % on cross validation of 10-fold. It exhibits classification accuracy of 78.3712% on test dataset which is more than all classifiers except C4.5 classifier. The experimental results also show that RMSE, true positive and false positive rates of proposed IDS are approximately same as C4.5, Naïve Bayes and Bagged Naïve Byes on Cross Validation. Moreover, RMSE, True Positive and False Positive rates of proposed IDS are better than all classifiers except C4.5 classifier on test dataset. Overall, the proposed intrusion detection system is very simple and accurate. The main disadvantage of proposed IDS is that it requires more time to build the model. So, onlinetraining of the IDS will is not be preferable.

Acknowledgements

We thank our Director Dr.L.M.Waghmare and Principal Dr. S.P.Danao for supporting us to write this paper. We also thank reviewers for improving the quality of the paper.

References

1. Eric Bauer and Ron Kohavi. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning, Kluwer Academic Publishers, Boston. Manufactured in The Netherlands 1998.
2. Thomas G.Dietterich. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning, Kluwer Academic Publishers, Boston 1999.
3. Shrinivasu and P.S.Avadhani. Genetic Algorithm based Weight Extraction Algorithm for Artificial Neural Network Classifier in intrusion Detection. In Procedia Engineering 38 (2012) 144 – 153, Published by Elsevier Ltd.,2012.
4. Li Hanguang, Ni Yu.Intrusion Detection Technology Research Based on Apriori Algorithm. In International Conference on Applied Physics and Industrial Engineering-2012.
5. Gisung Kim, Seungmin Lee and Sehun Kim. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. In journal of Expert Systems with Applications, Published by Elsevier-2014.
6. Wei Wang, Thomas Guyet, René Quiniou, Marie-Odile Cordier, Florent Masseglia and Xiangliang Zhang .Autonomic intrusion detection: Adaptively detecting anomalies over unlabeled audit data streams in computer networks. In Journal of Knowledge-Based Systems, published by Elsevier, 2014.
7. Wenying Feng, Qinglei Zhang, Gongzhu Hu and Jimmy Xiangji Huang. Mining network data for intrusion detection through combining SVMs with ant colony networks. In Journal Future Generation Computer Systems 37 127–140,2014.
8. Fangjun Kuanga, Weihong Xua and Siyang Zhang. A novel hybrid KPCA and SVM with GA model for intrusion detection. In Applied Soft Computing 18, 178–184,2014.
9. Abdulla Amin Aburomma and Mamun Bin Ibne Reaz. Evolution of Intrusion Detection Systems Based on Machine Learning Methods. In Australian Journal of Basic and Applied Sciences, 7(7): 799-813, ISSN 1991-8178.
10. Eibe Frank and Ian H. Witten. Generating Accurate Rule Sets Without Global Optimization. Department of Computer Science, University of Waikato, Hamilton, New Zealand.