BAGGING ENSEMBLE TECHNIQUE FOR INTRUSION DETECTION SYSTEM

Annkita Patel[1], Risha Tiwari[2]
Hasmukh Goswami College of Engineering, Ahmedabad, India

Abstract: Today almost no one can exclude himself or herself from using the Internet. Intrusion Detection system is software which helps us to protect our system from other system when other person tries to access our network. It secures our system resources without giving access to other system. This paper discusses intrusion detection systems built using ensemble techniques, i.e., by combining several machine learning algorithms. Network attacks can be divided into four classes: probe, remote to local, denial of service, and user to root. Experiments showed that Intrusion Detection Systems obtain better results when each class of attacks is treated as a separate problem by ensemble approach and handled by some specialized algorithms. In this paper ensemble technique bagging is applied on two different classifiers. Each module of the ensemble designed in this work is itself an ensemble created by using bagging of decision tree and support vector machine. Result obtain by combining this two classifiers is better than other ensemble technique.

Keywords: Intrusion Detection System (IDS), Bagging, decision tree and support vector machine (SVM).

I. INTRODUCTION

Today almost no one can exclude himself or herself from using the Internet. Like getting a cold or the flu once or twice a year, to Internet users, Internet attacks seem unavoidable. People take flu shots to prevent getting a virus and Internet users need protection to keep their network secure from attacks as well. Hence, intrusion detection systems [9] play an important role in modern network security. To design an intrusion detection system, a variety of techniques have been proposed over the past years. They are mainly categorized into two groups: anomaly detection techniques and misuse detection techniques. Misuse detection techniques model patterns of known attacks. By simply matching signatures of traffic records with previous well defined attack patterns, activities can be declared as intrusions if any mismatch happens. The recognized attacks are detected in an efficient way with a high level of accuracy. However, it is difficult to cover all possible variations of attacks using misuse detection technique because computer attacks are usually polymorphic. Because these polymorphic, novel attacks are constantly being introduced to the networks today it is necessary to be prepared to stop them before they do significant damage. For network intrusion detection system, a number of researchers say that the most powerful methods for extracting information hidden in large data sets, the data mining methods, implemented. Due to a large amount of processing required for network data, we can use data mining techniques [12]. To apply data mining techniques in intrusion detection, pre-processing data collected by the first step. Then used a special format for exchanging data mining process. After that, the configuration is used for classification and clustering. Rule based classification model, a decision tree-based, Bayesian network-based or based on the neural network. The data mining technology is used to ensure accuracy and efficiency in the search process, because any intrusion will not be missed while deal with a real time data.

II. BAGGING ALGORITHM

Bagging[1], which means bootstrap aggregation, is one of the simplest but most successful ensemble methods for improving unstable classification problems. In an ensemble using the bagging technique, all algorithms of the ensemble are used in parallel. In this case, each algorithm builds a different model of the data and the outputs of all predictors are combined to obtain the final output of the ensemble. In order to build different models, either each algorithm of the ensemble, or the data fed to each algorithm, or both, can be different. Since all algorithms perform in parallel, each of them can be executed on a different processor to speed up the computation. This is an important advantage over the boosting technique because nowadays multicore processors are very common even on personal computers. With this kind of architecture, the ensemble does not significantly increase the processing time compared to a single algorithm because the only additional time needed is used for the decision function that combines the outputs of all algorithms.

A. BAGGING ALGORITHM:

INPUT: S: training set; T: no of iterations;
n: bootstrap size
OUTPUT: BAGGED classifier: H(x) = majority (h1(x),...,hT(x)) where ∈ [-1, 1] are the induced classifier
Given training data (x1, y1),...,(xn, yn)
For t=1,..., T:

- form bootstrap replicate dataset St by selecting n random examples from the training set with replacement
- let ht be the result of training base learning algorithm on St output combined classifier: H(x) = majority(h1(x),...,hT(x))

III. VARIOUS METHODS FOR ENSEMBLE TECHNIQUE

We have analyzed various algorithms in ensemble technique as follows:

A. Ensemble of Machine Learning Algorithms for Intrusion

Detection [2]
In this paper, a three layer hierarchy multi-classifier intrusion detection architecture is proposed to promote the overall detection accuracy. For making every individual classifier is independent from others, each uses a diverse soft computing technique as well as different feature subset. In the kernel of base feature selecting classifiers they select a variety of supervised learning techniques that can provide capability of dealing with vagueness: fuzzy k-NN classifier, naive bayes classifier, and back propagation neural network classifier. In addition, the performances of a variety of combination methods that fuse the outputs from classifiers are studied. Having finished the process of base feature selecting classifiers' derivations in each group,, all the decisions from multiple ones are combined into a fused result. Finally, the predictions of three groups are then integrated to produce an ultimate conclusion of the ensemble. In order to evaluate the result of different combination methods, we carried out four fusion techniques: the majority voting rule, the average rule, Dempster-Shafer technique, and Bayesian combination method. In the experiments, DARPA KDD99 intrusion detection data set is chosen as the evaluation tools. The experimental results demonstrate that this hierarchy structure obtain a better detection performance than that of a single classifier using either partial feature subset or full feature set. The result also shows that the Bayesian combination method achieves the best detection accuracy among those four diverse combination techniques.

B. The Feature Selection and Intrusion Detection Problems [3]
They present in this paper an agent based IDS architecture that is capable of detecting probe attacks at the originating host and denial of service (DoS) attacks at the boundary controllers. They investigate and compare the performance of different classifiers implemented for intrusion detection purposes. Further, they study the performance of the classifiers in real-time detection of probes and DoS attacks, with respect to intrusion data collected on a real operating network that includes a variety of simulated attacks. Feature selection is as important for IDS as it is for many other modelling problems. They present several techniques for feature selection and compare their performance in the IDS application. It is demonstrated that, with appropriately chosen features, both probes and DoS attacks can be detected in real time or near real time at the originating host or at the boundary controllers. They also briefly present some encouraging recent results in detecting polymorphic and metamorphic malware with advanced static, signature-based scanning techniques.

C. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches [4].
In this paper, aim is to review the state of the art on ensemble techniques in the framework of imbalanced data-sets, with focus on two-class problems. They propose taxonomy for ensemble-based methods to address the class imbalance where each proposal can be categorized depending on the inner ensemble methodology in which it is based. In addition, they develop a thorough empirical comparison by the consideration of the most significant published approaches, within the families of the taxonomy proposed, to show whether any of them makes a difference. This comparison has shown the good behaviour of the simplest approaches which combine random under sampling techniques with bagging or boosting ensembles. In addition, the positive synergy between sampling techniques and bagging has stood out. Further- more, their results show empirically that ensemble-based algorithms are worthwhile since they outperform the mere use of preprocessing techniques before learning the classifier, therefore justifying the increase of complexity by means of a significant enhancement of the results. Finally, they have concluded that ensemble-based algorithms are worthwhile, improving the results that are obtained by the usage of data preprocessing techniques and training a single classifier. The use of more classifiers makes them more complex, but this growth is justified by the better results that can be assessed.

D. Cascading of C4.5 Decision Tree and Support Vector Machine for Rule Based Intrusion Detection System [5]
This paper represents two hybrid algorithms for developing the intrusion detection system. C4.5 decision tree and Support Vector Machine (SVM) are combined to maximize the accuracy, which is the advantage of C4.5 and diminish the wrong alarm rate which is the advantage of SVM. An intrusion detection system structure is proposed, and for estimating the performance, network data was required. Since the data collection for training and evaluating the classifier is a nontrivial task and there major was to promise the uprightness of the computer system, hence NSL KDD intrusion dataset is used for estimating the system. This system framework combines two classification algorithms as a core technique. After the testing is performed on NSL KDD dataset, numerical results demonstrate that there system have slight advantage over the KDD Cup 99. False alarm rate is low and high accuracy and less time is required by the proposed architecture. However attacks were not labeled, so there system only categorizes the connection as an abnormal or normal.

E. Ensembles of Decision Trees for Network Intrusion Detection Systems [6]
The aim of this work was to show that ensemble approaches fed with appropriate features sets can help tremendously in reducing both the number of false positives and false negatives. In particular, our work showed that the sets of relevant features are different for each class of attacks, which is why it is important to treat those classes separately. We developed our own IDS to evaluate the relevance of the sets of features. This system is an ensemble of four ensembles of decision trees. Each of the four ensembles is in charge of detecting one class of attacks and composed of four decision trees trained on different sets of features. The first three decision trees were fed with sets of five features. The last decision tree was fed with the union of these three sets of

five features from which the redundant features were removed. The experiments showed that these sets were appropriate in most cases. In the first experiment, the set of features selected by linear genetic programming gave the worst results, except for the class DoS for which the set of features selected by SVM performed poorly. The second experiment gave less interesting results because of the inappropriate distribution of examples between the training and test sets of the KDD99 data. Finally, a thorough analysis of the examples that were misclassified by the ensemble was performed, in particular highlighting the types of attacks that were systematically misclassified by the ensemble. By looking at the signatures of these attacks, we were able to find the reasons for the classification errors.

F. Cyber Security Threats Detection Using Ensemble Architecture [7]

This paper describes an ensemble design for cyber security threats detection, which fuses the results from multiple classifiers together to make a final assessment decision. For promoting both speed and accuracy in the detection performance, only some of the features in traffic data are selected for each base classifier. In the kernel of each classifier, they combine Dempster-Shafer theory with k-nearest neighbor technique to solve the uncertainty problems caused by ambiguous and limited intrusion information. In addition, they apply data mining techniques to reduce the number of false alarms. They apply the ensemble technique to our intrusion detection task and develop an ensemble feature selection approach, which includes six base classifiers that are created by diverse subsets of KDD99 data set. They have used three-layer hierarchy structure improves detection performance. The results indicate that their ensemble approach achieves higher detection rates than that of using a full feature set of classifiers.

IV. PROPOSED ALGORITHM

In this paper new proposed algorithm is presented which combine different classifier using bagging ensemble technique. The aim of the proposed system is to monitor and analyze network traffic in a computer network and collects network logs. Then the collected network logs are analyzed for feature selection by using data mining algorithms. Then ensemble technique is used. Here main motivation is to built a system using bagging ensemble technique with two different classifier.

Input : KDD 99 data set for intrusion detection

Begin :

Step 1 : Take the part of a KDD 99 data set as a training data.

Step 2 : Apply feature selection algorithm on training data for obtain better result. Here principle component method is used for feature selection.

Step 3 : In this step classification is done on the data set. First classify data by support vectore machiene(SVM) and then apply classification using decision tree(DT).

Step 4 : Combine the result of two classifier.

Step 5 : Apply ensemble technique bagging on the combine classifier .

This proposed algorithm reduces the problem of time requirement which is present in boosting ensemble technique. Also it gives higher accuracy by combining two classifiers with bagging ensemble technique. Result of this experiment is efficient and relevant to every user.

V. EXPERIMENTAL ANALYSIS

After doing the implementation of the existing approach, the comparison of the result is to be analyzed. After analysis we have seen that compare to bagging technique boosting gives more accuracy but at that time it also consumes more time than bagging. We have also seen that decision tree classifier gives more accuracy by correctly classified instant with compare to SVM but time taken to build model by decision tree is higher than SVM. So by combining both of the classifier we obtain better result. After this analysis we have done our proposed work with combine classifier. Result shows that bagging gives more accuracy than boosting and also bagging take less time than boosting.

Algorithm	Accuracy	Time
SVM+DT+Boosting	99.23%	41.06 sec
SVM+DT+Baggigng	99.89%	31.17 sec

VI. CONCLUSION

- In this research paper, we have implemented steps of our proposed algorithm that is preprocessign, classification and combine classifier.
- First we have done preprocessing step on data set using feature selection by principle component analysis.
- Then we have implemented classification using SVM and Decision tree with both ensemble technique: bagging and boosting. We have applied both techniques individually to the two different classifier and compare results.
- After comparison there is a clear picture about four different results.
- After individuale implementation we have combine both classifiers.
- Implementation of combine classifier using bagging ensemblle technique is done in the last step.
- After implementation we have seen that as shown in our proposed method bagging technique gives better result than boosting with combine classifier.

VII. ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude to my guide, Asst Prof. Risha Tiwari for her valuable guidance and useful suggestions.

REFERENCES

[1] L. Breiman, “Bagging predictors”, Machine Leaming, vol. 2, no. 24, (1996).
[2] Alexandre Balon-Perin, “Ensemble-based methods for intrusion detection”,2011-2012.

[3] Hui Zhao, "Intrusion Detection Ensemble Algorithm based on Bagging and Neighborhood Rough Set", International Journal of Security and Its Applications, 2013.
[4] A. H. Sung and S. Mukkamala, "The feature selection and intrusion detection problems," in Proceedings of the 9th Asian Conference on Advances in Computer Science. Chiang Mai,Thailand: Springer-Verlag, May 2004, pp. 468–482.
[5] Mikel Galar, Alberto Fern´andez, Edurne Barrenechea, Humberto Bustince, "A Review on Ensembles for the Class Imbalance Problem: Bagging, Boosting, and Hybrid-Based Approaches" IEEE transactions on systems, man, and cybernetics—part c: applications and reviews 2009.
[6] Jashan Koshal, Monark Bag, "Cascading of C4.5 Decision Tree and Support Vector Machine for Rule Based Intrusion Detection System" I.J. Computer Network and Information Security, 2012, 8, 8-20.
[7] Alexandre Balon-Perin and Bjorn Gamback, "Ensembles of Decision Trees for Network Intrusion Detection System" International Journal on Advances in Security, vol 6 no 1 & 2, year 2013.
[8] Heba F. Eid, Ashraf Darwish, Aboul Ella Hassanien, and Ajith Abraham, "Principle Components Analysis and Support Vector Machine based Intrusion Detection System" IEEE 2010.
[9] Te-Shun Chou, "Cyber Security Threats Detection Using Ensemble Architecture", International Journal of Security and Its Applications Vol. 5 No. 2, April, 2011.
[10] S. Mukkamala, A. Sung, and A. Abraham, "Cyber security challenges: Designing efficient intrusion detection systems and antivirus tools," in Enhancing Computer Security with Smart Technology, V. R. Vemuri and V. S. H. Rao, Eds. BocaRaton, Florida: CRC Press, Nov. 2005, pp. 125–161.
[11] Alexandre Balon-Perin, "Ensemble-based methods for intrusion detection",2011-2012.
[12] Weka: Data Mining Software in java http://www.cs.waikato.ac.nz/ml/weka/.
[13] KDD'99 dataset, (2010), http://kdd.ics.uci.edu/databases, Irvine, CA, USA.
[14] Jiawei Han , Micheline Kamber, (2003), " Data Mining – Concepts and Techniques" Elsevier Publications.
[15] en.wikipedia.org/wiki/
[16] A. Borji, "Combining Heterogeneous Classifiers for Network Intrusion Detection," Lecture Notes in Computer Science, Springer, Volume 4846, pp. 254-260, 2008.

