AN EFFICIENT ENSEMBLE MECHANISM FOR INTRUSION DETECTION

Abstract

Most classification and regression problems can be solved by training multiple learners, not by building a learner from the data. The boosting algorithm is one of the more recent developments in classification methodology, and it can turn weak learners into strong ones. It works by adding a classification algorithm to the updated weights of training samples in a sequential manner, through the majority voting technique of a sequence of classifiers. The AdaBoost algorithm is an efficient algorithm that combines weak algorithms to create a powerful classifier which can classify the training dataset with high accuracy. From the simulation results, it's clear that the AdaBoost classifier can achieve high search accuracy with short computation time and minimal cost, compared to the classification method. We have proposed a policy model to generate normal classes and attack classes and implemented an online access engine to allow or deny network access.

Keywords: Clustering, optimization, weak learners, strong learners, AdaBoost.IDS(Intrusion Detection System)genetic algorithm (GA), SVM (Support Vector Machine)

Authors

Dr. Pramod
Associate Professor & HOD
Department of CSD
PESITM
Shimoga, Karnataka, India
pramod74@pestrust.edu.in

Dr. Sunitha B S
Associate Professor & HOD
Department of CSDS
PESITM
Shimoga, Karnataka, India
sunitha.bs@pestrust.edu.in

Lohith C
Professor
Engineering Department
Garden City University
Bangalore, Karnataka, India

Sabin T.T
Assistant Professor
Department of ISE
SJCIT
Chickballapur, Karnataka, India
sabinsree@gmail.com

Divakar K.M
Assistant Professor
Department of CSE
SJCIT
Chickballapur, Karnataka, India
dnshgowda@gmail.com

I. INTRODUCTION

With significant advancements in the digital governance of e-commerce, social media, etc., the Internet is playing an increasingly significant role on a global scale. However, fear, criminal activity, and cyber attacks that started to build and launch highly complex attacks driven by destructive intentions have made the internet vulnerable. Our network resources and gadgets need to be secure, i.e., private, intact, and available [1] [2]. The process of locating and gathering information about hostile acts intended to jeopardize computer and network security is known as an intrusion detection system [1]. It is a crucial and delicate component of the deep security system, which also consists of intrusion detection, firewalls, program wrappers, and scanning and patching for vulnerabilities, access control, and encryption. Security will be needed to protect the Internet's vital infrastructure.

The main advantage of intrusion detection is the training and installation effort with the inference engine. We have antivirus and detection systems implemented in our network and we constantly strive to develop and carry out new attacks. As soon as information about a new attack is gathered by detection systems, it must be quickly integrated with current detection systems in order to prevent further damage from the new attack as soon as possible. However, due to training difficulties and large amounts of data, retraining models for existing and new attacks is often time-consuming. By the time new types of detection are available, new types of intruders may have done significant damage.

Intrusion is unauthorized access to hidden resources or restricted domains. This is how attackers gain unauthorized access to your network or private network. An attack is any suspicious or malicious activity on a network or computer. This is unauthorized interference with someone else's property. In retrospect, attackers try to identify security vulnerabilities before attacking her systems. To detect unwanted behavior that compromises security such as privacy, integrity, or availability. Amazing advances in technology and the Internet have created serious problems with computer security. Numerous machine learning, data mining and cognitive algorithms are the subject of modern advanced research aimed at improving diagnostics. His two types of detection methods are static detection (offline) and dynamic detection (online). It is a mechanism for instantly detecting various suspicious things on the network. Dynamic detection methods are effective, reliable, and efficient compared to static methods.

Learning Techniques: The process of creating a model from data is called learning or training. There are many learning styles and the two standard learning styles are supervised learning and unsupervised learning. In supervised learning, the goal is to estimate the target in real time. Unsupervised research does not rely on label information and aims to identify distributional information in the data. In other areas the coordinator is a single class or multiple coordinators. Multiple taxonomies are also called hierarchical taxonomies. During model design, feature selection and training techniques reduce computational cost, model size, range, and accuracy. In hierarchical systems, planners are often seen as weak and incapable of planning effectively. The search accuracy is close to zero. However, in many layers weak class systems can be combined into strong layers and defined with good detection rates [4]. Many learning algorithms are currently available for classifying samples or exemplars in datasets. Effective algorithms include linear discriminant analysis, neural networks, decision tree Navie-Bayes classifiers, nearest neighbors, and SVM [5].

The ensemble classification method is simply the process of combining many weak learner methods to create an efficient learner that can effectively organize the learning process. Good and efficient predictions can be generated if the settings are close or correct to the actual values [6]. Essentially, clustering algorithms are supervised algorithms because they can be trained to make their predictions true. You can find good ideas for making good predictions for a particular problem. The main idea of clustering methods is to use one central learner to generate many ideas [7].

II. LITERATURE SURVEY

The philosophy behind the classifier theorem is that a given classifier compensates for another classifier's error. However, training a classifier alone may not solve the desired problem because the classifier is not connected. Base manager is a distribution used to build a hierarchy. To build a classification system, we can consider a variety of weak learners such as support vector machines, nearest neighbors, and neural networks. Basic learning content is systematically created as extensions and containers. This upgrade adds weak learners and trains strong learners that provide better results and more accurate predictions. The boosting algorithm is based on Kearns and Valiant [1989] and his two complexity classes, the conceptual question of whether weak and hard learning problems are equal. This question is very important. Because any weak learner can become a strong learner if the answer is affirmative. In practice, however, it is often easy to find weak learners, but difficult to find strong learners. Schapire [1990] showed the answer in the affirmative, leading to the development of boosting algorithms [8].

With the rapid growth of the Internet and increasing global access to online content [16], the frequency of cybercrime has skyrocketed. Both end users and businesses are now vulnerable to cyber threats. It is important to put in place protective measures such as firewalls and IDSs. A firewall acts as an entry point, allowing or denying the passage of packets based on predefined criteria. In extreme cases, all network traffic may be blocked. Conversely, IDS automates the monitoring of computer networks. However, the continuous flow of data in such networks poses major challenges in developing effective IDSs, and a new approach utilizing online classification of datasets is proposed in this study to address this problem. Introduced in This method involves an incremental naive Bayesian classifier and uses active learning to achieve results on small sets of labeled data that are expensive to obtain. This approach involves his two sets of actions: offline, where data are preprocessed, and online, where the NADAL online method is introduced. A comparative analysis using the NSL-KDD standard dataset shows several advantages of the proposed method: (1) overcoming the challenges of streaming data, (2) the high cost associated with instance labeling. (3) improved accuracy and kappa values compared to the incremental naive Bayesian approach. Therefore, it turns out that this method is very suitable for IDS applications.

In our increasingly internet-reliant landscape, the main drawback of unauthorized intrusion into computer systems has escalated [17]. Intrusion refers to illicit access or activity within a computer tem. This highlights the growing importance of intrusion detection techniques to bolster overall computer system security. Intrusion detection involves identifying, preventing, or addressing intrusion attempts. This paper centers on an Intrusion Detection System driven by a GA. The technique applies GA to enhance network Intrusion

Detection Systems (IDSs). It provides an overview of IDT, genetic algorithms and related detection methods. The paper delves into GA parameters, evolution processes, and their intricate details. Notably, this implementation uniquely considers both temporal and spatial attributes of network connections when encoding connection information into IDS rules, which aids in identifying complex anomalous behaviors. The focus of this work lies in TCP/IP network protocols.

Intrusion Detection Systems can be classified intotwogroups based on their scope: Network-based IDS (NIDS) and Host-based IDS. NIDS observes intrusions by monitoring network traffic through devices like Network Interface Cards (NICs). Conversely, Host-based IDS monitors file and process activities within a specific host's software environment. Some host-based IDSs also analyze network traffic to detect attacks against a host.

A study [18] explored the impact of macro-level opportunity indicators on cyber theft victims and applied criminal opportunity theory to assess risk exposure. Use Internet access state patterns and other structural state characteristics to measure risk and estimate cyber damage impact. Furthermore, the percentage of users who access the Internet only from home is positively correlated with the prevalence of cyber theft. This study discusses the theoretical implications of these results. The role of link and node characteristics in social networks is investigated using the node classification method OS-ELM [19]. This method considers both node attributes and interactions for classification. Additionally, the Extreme Learning Machine (ELM) is used for intrusion detection within the IDS. Comparing the performance of SVM and ELM shows that the accuracy is comparable, but ELM has faster processing time. ELM is better than SVM in detecting intruders and the detection process takes less time.

III. PROPOSED METHODOLOGY

An ensemble learning process trains a large number of learners rather than individuals and combines the end results of various learners to achieve better results than weak individual learners. Therefore, it is also called a multiple classification system [6]. A collective process always consists of combining several ideas to create the best idea. In other words, organization is the activity of gathering a large number of weak learners for the purpose of producing strong learners. A set of words is often reserved for how multiple ideas can be derived in a single base learner. Broadly speaking, many classification systems also cover different ideas that do not come from a single elementary school student. Figure 1 shows a building block with n weak learners combined into her one strong learner. Weak learners, also called core learners, must be derived from a core learning algorithm (decision trees, neural networks, or any kind of learning algorithm). The main focus of combinatorial techniques is to combine the predictions of multiple models generated using learning techniques to improve the versatility or robustness of a single model.

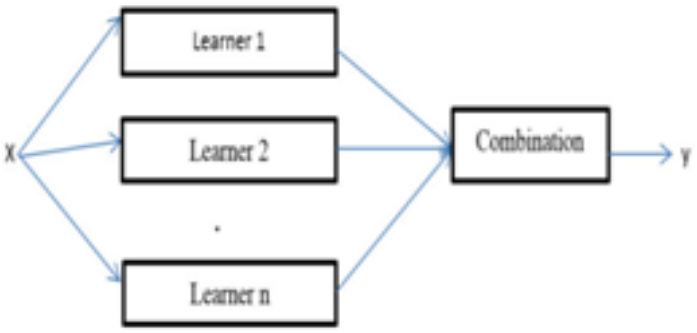

Figure 1: Simple ensemble architecture

There are basically two types of learners. Homogeneous learners and heterogeneous learners. The combinatorial method uses a special basic learning method that produces the same learners, i.e. learners who learn the same type, resulting in a unified system [8]. Different types of learners experience different processes and many learning algorithms are used. The generic skills of a particular group are often superior to the basic learner skills. In fact, combinatorial techniques can be of particular interest because they can turn weak learners into strong learners. Weak learners are good at making random predictions, while strong learners can make accurate predictions.

A. Proposed predictive model: An example of our proposed prediction is shown in Figure 1. Details of the model training and testing process are shown in Figure 2. 2. The training and testing process consists of his two parts, the online system and the offline system. In the offline method, the purpose of network traffic memory is to save training time, generate a function with the correct data structure, and match it with online network traffic behavior and corresponding model training. The proposed model consists of his three main components such as data processing, model training and inference engine.

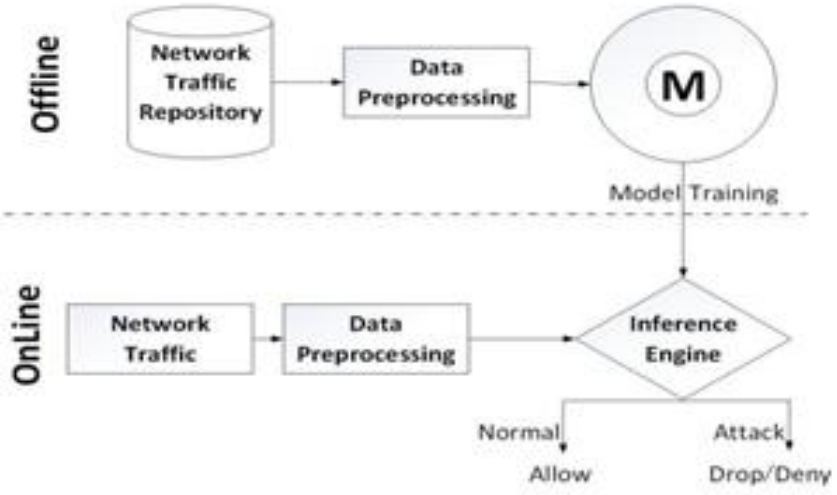

Figure 2: Train the model to classify traffic in network

Table 1: Confusion-Matrix for NSLKDD Dataset

Logitboost	PC	1	-1
AC	1	57993	637
	-1	450	66893

Table 2: Confusion-Matrix for KDD Corrected Dataset

Logitboost	PC	1	-1
AC	1	244476	655
	-1	5960	58938

Table 3: Performance Evaluation of the Proposed Model

DataSet	TPR	TNR	FNR	FPR	Accuracy	PPV	NPV	MCC	F1 Score
KDDCorrected	0.97	0.97	0.023	0.027	97.55	0.99	0.9	0.924	0.984
NSLKDD	.989	0.993	0.01	0.006	99.13	0.992	0.99	0.982	0.99

IV. RESULTS AND DISCUSSION

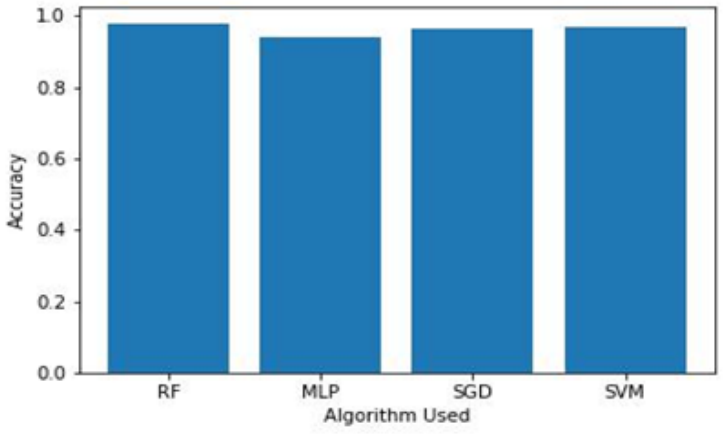

Figure 3: Accuracy Estimation of Existing Methods

Above graph tells the Accuracy of Existing Algorithms and its comparison.

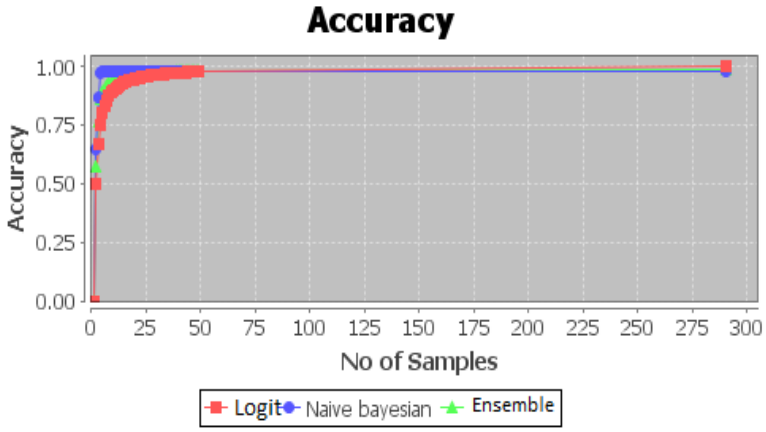

Figure 4: Accuracy Estimation of Proposed Methods

Above graph tells the Accuracy of Proposed Algorithms and its comparison.

V. CONCLUSION

Using clustering techniques to distinguish between regular traffic and congestion improves initial detection with minimal computation and cost compared to a single classifier. Ada Boost is an effective false positive technology to reduce false alarms. Using the same dataset for training and testing the proposed model yields high accuracy and low error. However, the accuracy is relatively low when using different datasets for training and testing. Combining three weak layers such as SVMs, neural networks, and decision trees can outperform individual layers. We conclude that adding more learners to the combinatorial model improves the recognition accuracy and reduces the probability of erroneous conditions occurring at each iteration.

REFERENCES

[1] Michael Kearns and Leslie Valiant. Cryptographic limitations on learning boolean formulae and finite automata. Journal of the ACM (JACM), 41(1):67–95, 1994.
[2] William Stallings. Network and internetwork security: principles and practice, volume 1. Prentice Hall Englewood Cliffs, 1995.
[3] Edward Amoroso. Intrusion detection: an introduction to internet surveillance, correlation, trace back, traps, and response. Intrusion. Net Book, 1999.
[4] Bernhard Scholkopf and Alexander J Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.
[5] Wray Buntine and Tim Niblett. A further comparison of splitting rules for decision-tree induction. Machine Learning, 8(1):75–85, 1992.
[6] Robert E Schapire. The strength of weak learnability. Machine learning, 5(2):197–227, 1990.
[7] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1):119–139, 1997.
[8] Yoav Freund and Robert E Schapire. Game theory, on-line prediction and boosting. In Proceedings of the ninth annual conference on Computational learning theory, pages 325–332. ACM, 1996.
[9] Erico N de Souza and Stan Matwin. Extending adaboost to iteratively vary its base classifiers. ´ In Advances in Artificial Intelligence, pages 384–389. Springer, 2011.
[10] Eric Bauer and Ron Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning, 36(1-2):105–139, 1999.
[11] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In Computational learning theory, pages 23–37. Springer, 1995.

[12] Giorgio Valentini and Thomas G Dietterich. Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. The Journal of Machine Learning Research, 5:725–775, 2004.
[13] Erin L Allwein, Robert E Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. The Journal of Machine Learning Research, 1:113–141, 2001.
[14] Simon Haykin and Neural Network. A comprehensive foundation. Neural Networks, 2(2004), 2004.
[15] David H Wolpert and William G Macready. No free lunch theorems for optimization. Evolutionary Computation, IEEE Transactions on, 1(1):67–82, 1997
[16] P. Alaei and F. Noorbehbahani, "Incremental anomaly-based intrusion detection system using limited labeled data," in *WebResearch (ICWR), 2017 3th International Conference on*, 2017, pp. 178–184
[17] Aman V. Mankar , Tushar C. Ravekar, "A Study of Intrusion Detection System using Advanced Genetic Algorithm" International Research Journal of Computer Science (IRJCS) ISSN: 2393-9842 Issue 11, Volume 3 (November 2016)
[18] H. Song, M. J. Lynch, and J. K. Cochran, "A macro-social exploratory analysis of the rate of interstate cyber-victimization,"*American Journal of Criminal Justice*, vol. 41, no. 3, pp. 583–601, 2016
[19] Sun, Y., Yuan, Y., & Wang, G. (2015). An on-line sequential learning method in social networks for node classification. *Neurocomputing*, *149*, 207–214. http://doi.org/10.1016/j.neucom.2014.04.074

