# An Efficient Ensemble mechanism for Intrusion Detection

Dr. Pramod1,

1Associate Professor, Department of ISE,

PESITM,

Shimoga, India

Email: pramod74@pestrust.edu.in

# Dr. Sunitha B S 2

2Associate Professor, Department of CSE,

PESITM,

Shimoga, India

Email: sunitha.bs@pestrust.edu.in

Lohith C3

3Assistant Professor, Departmentof CSE,

Gopalan College of Engineering and Management,

Bangalore, India

Email: lohithchandrashekhar@gmail.com

# Sabin T T4

4Assistant Professor, Department of ISE,

SJCIT,

Chickballapur, India

Email: sabinsree@gmail.com

# **ABSTRACT**

To solve classification or regression problems, ensemble methods or combination models train multiple learners, not by constructing a learner from the data but rather by constructing a set of learners and combining them together. One of the recent developments in classification methodology is the boosting algorithm. Boosting is a type of algorithm that can turn weak learners to strong learners. It works by adding a classification algorithm to the updated weight of training samples in a sequential manner. This is done through the majority voting technique of a sequence of classifiers. The boosting process combines the weak models to create a strong and reduce the noise of the combined model. The AdaBoost algorithm is an efficient algorithm that combines weak algorithms to create a powerful algorithm which can classify the training dataset. AdaBoost differs from existing optimization methods in terms of detection accuracy, error rate reduction, computation time and detection rate. Detection accuracy and computation cost are the two key metrics used to assess the execution of the AdaBoost classifier. From the simulation results, it is clear that the AdaBoost algorithm can achieve high search accuracy with short computation time and minimum 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)

## INTRODUCTION

The Internet plays an increasingly important role worldwide with great progress in the digital governance of e-commerce, social media, etc. But now, the internet is unsecured because of fear, criminal acts, cyber attacks that began to develop and launch very sophisticated attacks motivated by destructive goals. We ought to ensure security, i.e privacy, integrity and availability of our network resources and devices [1] [2]. Intrusion detection system is the process of identifying and gathering to malicious actions aimed at compromising computer and network security [1]. It is a critical and sensitive element of the deep security system, which includes: security policy, vulnerabilities, patching and scanning, access control and scanning, encryption, firewalls, program wrappers, and intrusion detection. Protecting critical infrastructure on the Internet will require security to prevent damage from invisible attack classes. Intrusion detection systems require efficient and accurate adaptation to acquire new knowledge and new classes that can be invented all the time [3]. This task can be solved by manually writing new attack methods and inserting them to existing systems that are distributed widely by automating learning new attack classes or adding them to deployed systems. .

The key benefit of intrusion detection are training and installation work using an inference engine. Anti-virus and detection systems are implemented in our network, will always try to develop and launch new attacks. Once information about a new attack is gathered through a detection system, it must have be quickly integrated into the current detection system to prevent further damage from a new attack as quickly as possible. However, retraining the model for already existing and new attacks is often slow due to difficulty of the training and the high volume data. By the time a new type of detection is ready, the new type of intruder might have caused signified damage.

Intrusion is illegal access to hidden resources or restricted domains. It is a way for attackers to gain access to a network or a private network without authorization. An attack is a suspicious or malicious activity on a network or computer. It is the act of trespassing on someone else's property. In hindsight, the attacker tries to identify weaknesses before attacking the security system. To detect unwanted behavior that compromises security such as privacy, integrity and availability. The huge increase in technology and the Internet is causing serious problems in computer security. Numerous machine learning, data mining and cognitive algorithms are the subject of modern and advanced research with the aim of improving the diagnosis. There are two types of detection methods such as static detection method (offline) and dynamic detection method (online). These are the mechanisms that are used to immediately detect various suspicious things on the network. Dynamic detection methods are effective, reliable and efficient compared to static methods

Learning Techniques: The process of creating models from data is called learning or training. There are numerous learning styles and two standard learning styles are supervised learning and unsupervised learning. In supervised learning, the goal is to estimate the target in real time. An unsupervised study do not depend upon on label information and aims to identify some distributional information in the data. In other areas, the coordinators are a single class or multiple coordinators. Multiple classifications are also called hierarchical classifications. During model design, feature selection and learning techniques lead to reductions in computational cost, model size, scope, and accuracy. In a hierarchical system, the scheduler is sometimes treated as weak and unable to schedule effectively. The accuracy of the search is close to zero. But in many levels, the system of weak classes is combined to form a strong level, which might be defined with a good detection rate [4]. Many learning-algorithms are available currently for classification of samples or samples in a dataset. Some effective algorithms are linear discriminant analysis, neural network, decision tree Navie Bayes classifier, nearest neighbor and SVM [5]

Ensemble Classification Methods is nothing but a process of combining the methods of many weak learners together to create a efficient learner who can organize the learning process effectively. It can produce a good and efficient forecast when its configuration is close to the real value or correct [6]. Basically, the clustering algorithm is a supervised based algorithm because it can train and make the prediction true. For a given problem, it can find the right idea to make a good prediction. The main idea behind the clustering method is to use one central learner to generate many ideas [7].

## LITERATURE SURVEY

The philosophy behind the classifier set is that the set classifier compensates for errors made by another. But just training a classifier may not solve the desired problem because classifiers are not connected. Base managers are the distributions that are used to build the structure of the hierarchy. We can consider different weak learners such as support vector machine, nearest neighbor and neural network to build classification system. Basic learners are created systematically as enhancements and similarly as containers. By upgrading, weak learners are added to train a strong learner to provide better results and accurate predictions. The boosting algorithm in accordance with conceptual question of Kearns and Valiant [1989], whether two classes of complexity, weak learning problems and hard learning problems are equal. This question is very important because if the answer is positive, any weak learner can become a strong learner. But in actual practice, it is often easy to find weak learners but difficult to find strong learners. Schapire [1990] demonstrated that the answer was affirmative, which led towards the development of a boosting algorithm [8].

With the rapid growth of the internet and increased global accessibility to online content [16], the incidence of cybercrime has surged. Today, both end users and enterprises face vulnerabilities to cyber threats. Employing defensive measures such as firewalls and IDS has become crucial. A firewall operates as entry point, allowing or denying the passage of packets based on predefined criteria. In extreme cases, it can even block all network traffic. Conversely, an IDS automates the surveillance of computer networks. However, the continuous flow of data in such networks presents a substantial challenge in developing effective IDS. To address this, a novel approach is introduced in this study, utilizing online classification of datasets. The method involves an incremental naive Bayesian classifier and leverages active learning to achieve results with a small set of labeled data, which can be costly to obtain. The approach encompasses two sets of actions: offline, involving data preprocessing, and online, introducing the NADAL online method. Comparative analysis using the NSL-KDD standard dataset demonstrates several advantages of the proposed method: (1) surmounting the streaming data challenge, (2) mitigating the high cost associated with labeling instances, and (3) enhancing accuracy and Kappa scores compared to the incremental naive Bayesian approach. Thus, the method proves to be highly 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 intotwo groups 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.

Examining the impact of macro-level opportunity indicators on cyber theft victimization, a study [18] applies criminal opportunity theory to assess risk exposure. It measures risk through state-level patterns of internet access, alongside other structural state characteristics, to gauge their influence on cyber-victimization. Furthermore, the proportion of users accessing the internet solely from home is positively correlated with high-level cyber theft victimization counts. The study discusses the theoretical implications of these findings. The role of links and node features in social networks is explored using the OS-ELM [19] node classification technique. This method considers both node attributes and interactions for classification. Additionally, the Extreme Learning Machine (ELM) is applied to intrusion detection within IDS. Performance comparisons between SVM and ELM indicate comparable accuracy, but ELM exhibits faster processing times. ELM outperforms SVM in detecting intrusions, requiring less time for the detection process.

## PROPOSED METHODOLOGY

The ensemble learning process trains many learners instead of one individual and combines the end results of different learners to achieve better outcomes than individual weak learners. Hence, it is also referred multiple classification system [6]. The collective process is always combining multiple ideas to create the best idea. In other words, organization is a activity to combine a vast number of weak learners with the goal of creating a strong learner. A set of words is often reserved for methods that results in multiple ideas using a single base learner. Broadly speaking, many classification systems also cover a hybrid of ideas that do not come from a single primary learner. Figure 1 shows a building block where n number of weak learners are combined to form a strong learner. Weak learners are also called as core learners which are need to be derived from core learning algorithms which can be decision trees, neural networks or any type of learning algorithm. The main focus of the combination method is to combine the predictions of several models that are build using learning methods to enhance the general or robustness of a single model.

##

**Figure 1: Simple ensemble architecture**

Basically, there exist two types of learners in general: homogeneous learners and heterogeneous learners. The combination method uses a special foundational learning method that creates learners with the same, that is, those one who learn the similar type, leading to a unified system [8]. If the learners are different types that lead to different processes, and use many learning algorithms. The general ability of a particular set is often stronger than that of basic learners. In fact, combination methods can be mainly interesting because they can improve weak learners to become strong learners. Weak learners are better at random predictions, but strong learners can make accurate predictions.

1. **Proposed predictive model**

An example of our proposed prediction is given in fig. 2 and particluars of the model training and testing process is shown in fig. 2. The training and testing process has two parts, comparatively, the online system and the offline system. In the offline method, the network traffic memory is there to save the training time and generate the features using the correct data structure and match that with the online network traffic behavior and the appropriate model training. . The Proposed model has three key components such as data processing, model training and inference engine



**Figure 2: Model training to classify the network traffic**

**Table 1: Confusion Matrix for NSLKDD Dataset**

****

**Table 2: Confusion Matrix for KDD Corrected Dataset**

****

**Table 3: Performance Evaluation of the proposed model**



## 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.

## CONCLUSION

Distinguishing normal traffic from congestion using clustering method improves the original detection with minimum computation time and cost compared to single classifier. AdaBoost is an effective false detection technique to reduce false alarms. For the proposed model, when we use the same dataset for training and testing, the correct percentage is high and the error is small. But with distinct data sets for training and testing, the accuracy rate is relatively low. Three weak layers such as SVM, neural network and decision tree ,which are combined and their performance is better than that of individuals. We conclude that by adding more learners to the combination model, the detection accuracy increases and the probability of incorrect conditions decreases in 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>