ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

Survey on Enhanced Anomaly Based Intrusion Detection System Using Limited Labeled Data

Pradnya Patil

Assistant Professor, Department of Computer Science and Engineering, TKIET Waranangar, Maharashtra, India

ABSTRACT: In recent era the network dramatically extended, security considered as major issue in networks. Internet attacks are increasing, and there have been various attack methods, consequently. Intrusion Detection System (IDS) is an effective security tool that helps to prevent unauthorized access to network resources by analyzing the network traffic and classifying the record as either normal or anomalous in this paper proposed method but has significant challenges in building IDS that are 1) Streaming nature of data and computer networks, 2) Feature selection or reduction because feature may be irrelevant or redundant and may inhabit system performance [1].
The proposed method includes online and offline classification on data set. For this Naive Bayes Classifier is used [2], after that active learning enables to solve the problem using subset of labeled data points. Here, we introduced the Network Anomaly Detection Using Active Learning (NADAL) online method that allows us three advantages, 1) Overcoming streaming data challenges, 2) Reduce the high cost with instance labeling, 3) improved speed detect 4) Accuracy of detection

KEYWORDS: Intrusion Detection, Anomaly Detection, Active Learning, Accurate identification

I. INTRODUCTION

In recent years, the rapid growth of network-based services and technologies has resulted in a surge in the number of network-based computer attacks. An *attack* refers to a set of actions that compromise the confidentiality, integrity, and accessibility of resources. A system is known to be secure if it can guarantee these three criteria. Attacks must be identified before doing any harm to the organization. Even Local Area Networks (LAN) need to be able to withstand such attacks since network performance is important in terms of bandwidth and other resources. The most common means of defense against potential attacks involves a two-layered system. The first layer comprises a firewall which controls access to the network while the second layer is configured to detect threats that somehow manage to pass through the firewall and take appropriate action to defend the network. This second layer is known as an **Intrusion Detection System (IDS)** which is able to identify intrusion attempts by monitoring and analyzing network packets and logs. In case an intrusion is detected, the system alerts the network administer [1-3].
Intrusion Detection System (IDS) is meant to be a software application which monitors the network or system activities and finds if any malicious operations occur. Tremendous growth and usage of internet raises concerns about how to protect and communicate the digital information in a safe manner. Nowadays, hackers use different types of attacks for getting the valuable information. Many intrusion detection techniques, methods and algorithms help to detect these attacks.

The main challenge is that attackers are always keeping novelty in their tools and techniques in exploiting any kind of vulnerabilities. Hence, it is very difficult to detect all types of attacks based on single fixed solutions. For that intrusion detection system (IDS) became an essential part of network security. It is implemented to monitor network traffic in order to generate alerts when any attacks appear. IDS can be implemented to monitor network traffic of a specific device (host intrusion detection system) or to monitor all network traffics (network intrusion detection system) which is the common type used.

An Intrusion Detection System (IDS) is a software application that is used to inspect the activities such as suspicious behavior, system attack, and misuse in a system. The Anomaly and Misuse detection are two main

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

categories of IDS techniques. Misuse Detection is based on signatures for known attacks and Anomaly detection can detect unknown attacks, but has high false positive rate. Due to this reason, it becomes an essential thing in Network Intrusion Detection Systems (NIDS). Most anomaly based NIDSs make use of supervised algorithms. The performances of these algorithms highly depend on training data but in real world it is difficult to obtain such training data in network environment. Apart from this, as increase network services also a challenge [5] In general, there are two types of IDS (anomaly base or misuse base). Anomaly intrusion detection system implemented to detect attacks based on recorded normal behavior. Therefore, it compares the current real time traffics with previous recorded normal real time traffics, this type of intrusion detection system is widely used because it has the ability to detect the new type of intrusions. But from another perspective, it registers the largest values of false positive alarm, which means there is a large number of normal packets considered as attacks packets. However, misuse intrusion detection system is implemented to detect attacks based on repository of attacks signatures. It has no false alarm but at the same time, the new type of attack (new signature) can succeed to pass-through it.

II. RELATED WORK

In order to meet the challenge to classify large volume of data, some efficient algorithms such as decision tree, naïve Bayesian [6], neural network, support vector machine [7], knearest neighbors, fuzzy logic model [8], and genetic algorithm [9] with necessary processing have been used widely in the last decades. In [10], Axellson writes a well-known paper that uses the Bayesian rule of conditional probability. In [11], a behavior model is introduced that use Bayesian techniques to obtain model parameters with maximal a-posteriori probabilities. In [12], author uses 3 feature ranking algorithms (Support Vector Machine, Multivariate Adaptive Regression Splines, and linear Genetic program) to reduce the feature space. In [13] author investigates the performance of two feature selection algorithm with an hybrid architecture by combining Bayesian network and Classification & Regression Tee. In [14], a dynamic model "Intelligent Intrusion Detection System" proposes for intrusion detection (anomaly, misuse and host based detection) where fuzzy neural networks is used. In 2007, Mrutyunjaya Panda et al. detect attack categories as well as intrusion by using Naïve Bayes classifier and back propagation neural network (BPN) and show the comparison between them where NBC performed better at 95% for 65,525 samples of KDD'99 dataset than BPN [15]. In [16], Prashanth et al. propose a two phases approach in intrusion detection design where the first phase select features and the second phase design a method to solve uncertainty problem at routers by using Random Forest algorithm in 2008 but the execution time increases with the increase of trees [17]. In [18], authors propose Enhanced Support Vector Decision Function for feature selection. In [19], authors examine an automatic feature selection procedure named Correlation– based Feature Selection. To detect network intrusion, Dartiguet et al. empirically analysis C4.5 binary classification based algorithm in 2009, and compare with other previous works by reducing bias, and bias & variance. Although this classifier perform well for normal attacks, but the overall accuracy is varies from around 70% to 94% with respect to information ratio and information Gain ratio [1]. In [20] Abdulla et al. show, the performance of IDS may be improved through selecting right features. Dewan et al. improve the performance compared to other previous researches by combining NBC and Decision Tree in 2010, where 494020 samples are exempted and get the overall accuracy at 90% [21]. Iftikhar et al. uses PCA to select features corresponding to the highest Eigen values using Genetic Algorithm [22]. A Feature reduction process named Feature Vitality Based Reduction Method (FVBRM) uses in Intrusion Detection method with NBC by Saurabh et al. in 2012 [23]. There use some other methods too but FVBRM evaluated 97.8% overall which is better than other methods. In 2015, Zhang et al. develop a new system named Hadoop-Based System to detect Website Intrusion and get overall accuracy around 98.5% by using hybrid IDS with random forest algorithm. All of these mentioned experiments use KDD'99 dataset but someone analysis by using 10% data of that dataset and someone analysis on less than 10% [24].

In [1], an online Bayesian classifier is constructed which distinguishes between normal and intrusive links in the KDD- 99 dataset. The classifier starts with a small number of training instances of both normal and intrusive classes. The remaining instances are then classified while the mean and standard deviation of the features are continuously updated. A key action in this online naive Bayesian classifier is to update the μ and σ values following each instance test. The method carries out naive incremental updating. Many modern intrusion detection methods focus on feature selection or reduction. This is because many features may be irrelevant or redundant and may inhibit system performance. In [11], important features are identified through reduced input.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

III. PROPOSED WORK

Naive Bayesian Classification

Naive Bayesian classification is a popular method for stream mining. The popularity of the method is due to the fact that the model can be updated with new data streams very easily. The method is inherently incremental since new data points are updated as they arrive. Given this incremental nature, the algorithm is very suitable to stream mining [12].

Naive Bayes Classifier: this classifier refers to the group of probabilistic classifiers. It implements Bayes theorem for classification problems. The first step of Naive Bayes classifier is to determine the total number of classes (outputs) and calculate the conditional probability for each dataset classes. After that the conditional probability would be calculated for each attribute. The standard formula of Naive Bayes can be found in the referred research [8]. Furthermore, it has the ability to work with discreet and continuous attributes also on the contrary of MLP classifier Naive Bayes can be implemented within a short period of time [11]. Meanwhile Naive Bayes can be represented as Bayesian network (BN) or Belief network. BN supports presenting independent conditional probability based on understanding framework. In general BN is acyclic graph between expected class (output) and a number of attributes [20] [25].

Active Learning Instead of inquiring about the correct labels for all instances, active learning determines how input instances are selectively labeled. Quite often, this approach requires considerably fewer instances to learn a concept, compared to typical supervised methods. The majority of research on the topic is focused on tuple selection for labeling.

In active learning, once an instance is scanned, depending on the selected strategy, the algorithm searches for the correct label and the predictive model is trained with the new instance. In the following, we briefly explain four active learning strategies • Random Strategy: Input samples are given random labels. • Fixed Uncertainty Strategy: The instances for which the current classifier has minimum confidence are labeled. A constant threshold is considered. Only those instances are labeled for which the maximum posterior probability as estimated by the classifier does not exceed the threshold. • Variable Uncertainty Strategy: Instances below the threshold are labeled with a time interval; the threshold is introduced as varying with time; and the budget is spent in a uniform fashion over time. • Uncertainty Strategy with Randomization: A random threshold is selected and the labels for instances near the threshold are inquired. [28]

Feature Selection or reduction:

To select the particular feature large data is available in the network and they are usually evaluated for intrusion. For example, the Internet Protocol (IP) address of the source and target system, protocol type, header length and size could be taken as a key for intrusion Feature selection is an effective and a vital element in high dimensionality data mining. It is a priori data processing steps in learning algorithm. Several feature subset selection techniques have used in data mining. [26]Feature reduction may use three standard feature selection methods: correlation, information gain, or gain ratio. The proposed method in this study employs feature vitality based reduction. The results indicate that the proposed model provides better performance.

The proposed model, called Network Anomaly Detection using Active Learning (NADAL) involves an offline and an online step. The selected dataset is preprocessed in an offline fashion. The NSL-KDD dataset contains instances labeled with the attack type. During the preprocessing step, the attacks are divided into four categories: DoS, Probe, R2L, and U2R. Furthermore, there are four classifiers at the respective layers of attacks. Thus, the preprocessing carried out using Weka selects the appropriate features for each classifier. The selected features are then given to the feature filtering module in NADAL. Figure 1 illustrates the NADAL framework. In the proposed online method, at each time, each instance is processed at most once to improve the model. The instance is then discarded. Initially, instance Xt having label yt passes through the feature filtering module and the appropriate features for each classifier are considered. At each layer, the naive Bayesian module incrementally predicts the probability that the instance belongs to the class. Thereafter, the selected active learning strategy (i.e. uncertainty with randomization) is called. The output of the strategy determines whether the label for the instance must be inquired. A logical OR gate is used to aggregate the results from different active learning modules. The classifiers are updated using the instance if the gate outputs 1. Otherwise, the aggregate output module predicts the label according to the maximum certainty calculated by the classifiers. In this case, Yt represents the actual label for instance Xt.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

Evaluation Criteria details:
The results are evaluated according to accuracy and Kappa. Accuracy represents the percentage of tuples in the dataset that are correctly labeled. The measure is calculated as below:

$$accuracy = \frac{TP + TN}{P + N}$$

The Kappa coefficient measures the agreement among individuals who classify or measure items. The value is obtained as follows:

$$Kapaa = \frac{p_o - p_c}{1 - p_c}$$

Where P_0 and P_1 denote observed and chance agreement, respectively.

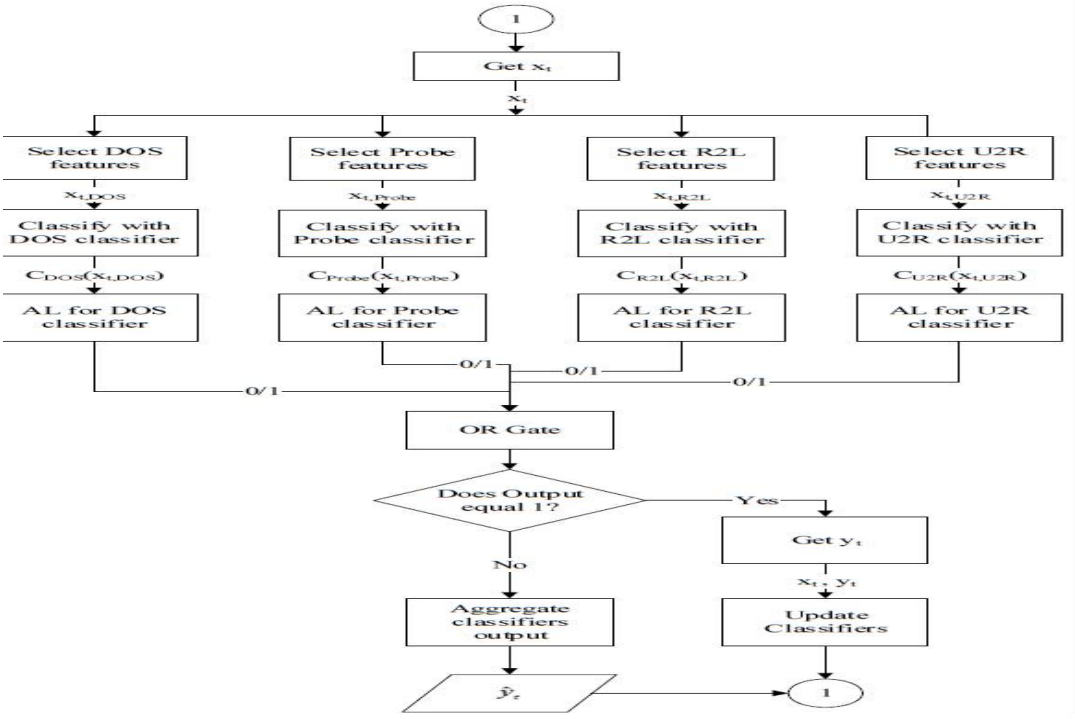

Fig. 1. The proposed model called NADAL
Proposed model of NADAL et.al [1]

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

IV. CONCLUSION AND RECOMMENDATION

Traditional data packets are inherently static. In contrast, streaming data are continuously created; they cannot be stored; and must by analyzed as a single unit. In this paper, a novel network anomaly detection framework was proposed to improve efficiency in classifying data in an online fashion. Furthermore, active learning was used to reduce labeling costs. The proposed system was evaluated using the standard NSL-KDD dataset. Implementation results revealed that the proposed method outperforms the naive Bayesian approach in terms of both accuracy and Kappa. There are many challenges in detecting network anomalies which can be addressed in future studies.

Our recommendations are as follows: • Employing other incremental classification approaches in NADAL and comparing the evaluation criteria. • Improving classification accuracy in data with class imbalance so that the data are equally distributed among the training classes. • Detecting concept drift in data streams where the relationship between input data and labels may be modified due to concept drift. The

REFERENCES

[1] Parisa Alaei , Fakhroddin Noorbehbahani, "Incremental Anomaly-based Intrusion Detection System Using Limited Labeled Data", 2017 3th International Conference on Web Research (ICWR)
[2] F. Gumus, C. O. Sakar, Z. Erdem, and O. Kursun, "Online Naive Bayes classification for network intrusion detection," in Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014, pp. 670–674.
[3] A. Rasoulifard, A. Ghaemi Bafghi, and M. Kahani, "Incremental hybrid intrusion detection using ensemble of weak classifiers," Commun. Comput. Inf. Sci., vol. 6 CCIS, pp. 577–584, 2008.
[4] R. Singh, H. Kumar, and R. K. Singla, "An intrusion detection system using network traffic profiling and online sequential extreme learning machine," Expert Syst. Appl., vol. 42, no. 22, pp. 8609–8624, 2015.
[5] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," Comput. Secur., vol. 28, no. 1–2, pp. 18–28, 2009.
[6] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: methods, systems and tools," IEEE Commun. Surv. Tutorials, vol. 16, no. 1, pp. 303–336, 2014.
[7] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Survey on incremental approaches for network anomaly detection," arXiv Prepr. arXiv1211.4493, 2012.
[8] Mukkamala, Srinivas, Guadalupe Janoski, and Andrew Sung. "Intrusion detection using neural networks and support vector machines." Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on. Vol. 2. IEEE, 2002.
[9] Luo, Jianxiong, and Susan M. Bridges. "Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection." International Journal of Intelligent Systems 15.8 (2000): 687-703.
[10] Yu, Yan, and Hao Huang. "Ensemble approach to intrusion detection based on improved multi-objective genetic algorithm." Ruan Jian Xue Bao(Journal of Software) 18.6 (2007): 1369-1378.
[11] Axelsson, Stefan. "The base-rate fallacy and its implications for the difficulty of intrusion detection." Proceedings of the 6th ACM Conference on Computer and Communications Security. ACM, 1999.
[12] Puttini, Ricardo S., Zakia Marrakchi, and Ludovic Mé. "A Bayesian classification model for real-time intrusion detection." AIP Conference Proceedings. IOP INSTITUTE OF PHYSICS PUBLISHING LTD, 2003.
[13] Sung, Andrew H., and Srinivas Mukkamala. "The feature selection and intrusion detection problems." Advances in Computer Science-ASIAN 2004. Higher-Level Decision Making. Springer Berlin Heidelberg, 2004. 468-482.
[14] Chebrolu, Srilatha, Ajith Abraham, and Johnson P. Thomas. "Feature deduction and ensemble design of intrusion detection systems." Computers & Security 24.4 (2005): 295-307.
[15] Bashah, Norbik, Idris Bharanidharan Shanmugam, and Abdul Manan Ahmed. "Hybrid intelligent intrusion detection system." World Academy of Science, Engineering and Technology 11 (2005): 23-26.
[16] Panda, Mrutyunjaya, and Manas Ranjan Patra. "Network intrusion detection using naive bayes." International journal of computer science and network security 7.12 (2007): 258-263.