Azerbaijan Journal of High Performance Computing, Vol 3, Issue 2, 2020, pp. 196-206
https://doi.org/10.32010/26166127.2020.3.2.196.206

Anomaly Detection Using Machine Learning Approaches

Mausumi Das Nath, Tapalina Bhattasali

St. Xavier's College (Autonomous), Kolkata, India, m.dasnath@sxccal.edu, tapalina@sxccal.edu

*Correspondence:
Mausumi Das Nath,
St. Xavier's College
(Autonomous), Kolkata,
India, m.dasnath@sxccal.edu

Abstract

Due to the enormous usage of the Internet, users share resources and exchange voluminous amounts of data. This increases the high risk of data theft and other types of attacks. Network security plays a vital role in protecting the electronic exchange of data and attempts to avoid disruption concerning finances or disrupted services due to the unknown proliferations in the network. Many Intrusion Detection Systems (IDS) are commonly used to detect such unknown attacks and unauthorized access in a network. Many approaches have been put forward by the researchers which showed satisfactory results in intrusion detection systems significantly which ranged from various traditional approaches to Artificial Intelligence (AI) based approaches.AI based techniques have gained an edge over other statistical techniques in the research community due to its enormous benefits. Procedures can be designed to display behavior learned from previous experiences. Machine learning algorithms are used to analyze the abnormal instances in a particular network. Supervised learning is essential in terms of training and analyzing the abnormal behavior in a network. In this paper, we propose a model of Naïve Bayes and SVM (Support Vector Machine) to detect anomalies and an ensemble approach to solve the weaknesses and to remove the poor detection results.

Keyword: Naïve Bayes, SVM, Hybrid Classifier, Ensemble, Anomaly Detection.

1. Introduction

The voluminous amount of data daily transmitted due to the enormous usage of the Internet. Malicious activities are increasing day by day in the network environment. Damage caused due to network attacks may vary from a little disruption in service to developing colossal loss. To monitor malicious activities, an efficient intrusion detector needs to be designed. Misuse or Signature detection approaches used in Intrusion detection can monitor network traffic to detect known attacks. An anomaly detection method can be used to flag deviation from normal usage patterns as an intrusion to deal with unknown attacks. Automated detection is preferable to deal with unknown malicious activities. Machine Learning approaches are well-suited for automated detection. A network anomaly is a sudden and short-lived deviation from the normal operation of the network. Machine learning algorithms can learn from data and make predictions based on that data. Anomaly detection is the procedure to identify items or events that do not conform to the expected pattern or to other items present in a dataset. Some anomalies are deliberately caused by intruders with malicious intent, while others may be purely an accident. Quick detection is needed to initiate a timely response, such as raising the alarm if a surveillance network detects an intruder. Network monitoring devices collect data at high rates.

Consequently, designing an effective anomaly detection system involves extracting relevant information from a voluminous amount of noisy, high-dimensional data. As different anomalies may exhibit themselves in different manners, developing a general anomaly detector model is difficult. Non-parametric learning algorithms based on machine learning principles are desirable as they can learn the nature of normal measurements and autonomously adapt to variations in the structure of "normality". Various statistical and machine learning approaches have been implemented to detect intrusion and protect the network. The machine learning approach is more comfortable as it can characterize what is expected in data using a simple mathematical model. It is useful to learn the characteristics of a system from observed data. Feature Selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for three reasons:

- Simplification of models to make them easier to interpret
- Shorter training times
- Enhanced generalization by reducing overfitting

Among the machine learning approaches, supervised learning, and unsupervised learning techniques have yielded satisfactory signatures and anomaly detection. Moreover, it can customize the normal activity for every system, network, and application. This increases the attacker's difficulty level to know or predict what can be done without getting detected. Among the various machine learning approaches used to handle anomaly detection, the supervised learning technique has outperformed the unsupervised learning approach. Furthermore, among the different supervised learning techniques, the Support vector machine (SVM) yields better results because of its characteristics - finding the optimal separation hyperplane, dealing with high dimensional data, and achieving accuracy in anomaly detection. Researchers have combined various techniques with SVM to offer anomaly detection efficiently and effectively in a networked environment. An anomaly detector model with an ensemble approach is considered here that integrates Naive Bayes logic and SVM to detect and retrieve network anomalies. The remaining paper is organized as follows. In Section 2, the literature survey of anomaly intrusion detection systems has been written. Section 3 formulates the problem statement, and section 4 briefly specifies the primary objective to carry on this research work. Section 5 illustrates our proposed work. Section 6 briefly analyses the proposed work. Finally, section 7 concludes this paper.

2. Literature Survey

An Outlier Detection technique was introduced by Jabez and Muthukumar (Jabez, J., & Muthukumar, B., 2015) to detect the proliferation in the computer network, which has been tested with the KDD datasets. The results yielded were better than other existing machine learning approaches. Sinclair et al. (1999) described how the various machine learning methods were used to create policies to detect an intrusion in the network system. On the other hand, Omar et al. (2013) depicted that supervised learning like multi-layer perceptron, SVM, and the rule-based methods yielded better results than other techniques. However, in unsupervised approaches, K-Means, SOM, and 1-class SVM achieved better results than the other existing approaches. Zhong and Khoshgoftaar (2007) carried out an experimental analysis of multiple centroid-based unsupervised clustering algorithms. They proposed an effective self-labeling heuristic but simple method to detect an attack and a normal group of network traffic audit data that might hinder the system. Moreover, it has been compared and seen that clustering-based methods work better in identifying unseen or new attacks. Cluster-based intrusion detection architecture was carried out by Sen (2010) for wireless ad hoc networks. The results obtained proved its efficiency as well as its effectiveness. A hybrid feature selection method using Correlation-based Feature Selection and Information gain was proposed by Wahba et al. (2015). Small numbers of features were used applying Adaptive boosting using Naive Bayes to improve the detection accuracy. Results showed that the proposed technique achieved higher accuracy in detecting attacks compared to other methods. On the other hand, Almseidin et al. (2017) carried out several experiments on the KDD intrusion dataset to evaluate and assess various machine learning classifiers. The implemented experiments yielded that the decision table classifier stated that the false-negative was very low, while the random forest classifier suggested a higher accuracy rate. Reviews were carried out by Tsai et al. (2009), which showed that combining ensemble and hybrid classifiers for intrusion detection proved to be a good one in terms of detection rate. Ren et al. (2019) used a hybrid data optimization technique to detect anomalous behavior in a network. Random Forest was used to building an intrusion detection system utilizing the most favorable training dataset achieved by data sampling and the feature selection method's selected features. The paradigm proved to have had several benefits over other methods. Haq et al. (2015) focused on the architectural issues of single, hybrid, and ensemble classifier design. It guided the researchers that hybrid or ensemble classifiers will perform better in several ways. A density-based and grid-based clustering algorithm was proposed by Leung and Leckie (2015). 1999 KDD Cup data set was used for estimation. The accuracy and computational complexity were noteworthy. Again, Wang et al. (2010) proposed a new approach, termed as FC-ANN, based on Artificial Neural Network and fuzzy clustering, which helped IDS achieve less false positive rate, higher detection rate. It outperformed well-known methods like decision tree, Naïve Bayes concerning detection accuracy and was more stable. Vinchurkar and Reshamwala (2012) discussed various machine learning approaches, along with dimension reduction using PCA. The present challenges were discussed, and it also stated how the system should be designed so that threats and security attacks could be detected. Further, Esmaily et al. (2015) proposed a combination of the Multi-Layer Perceptron (MLP) ANN and Decision Tree (DT) algorithm to identify attacks accurately and reliably. Sabhnani and Serpen (2003) depicted pattern recognition and machine learning approaches using the KDD 1999 Cup intrusion detection dataset to identify four different types of attacks. The results obtained were a significant improvement over other approaches. Belavagi and Muniyal (2016) furthermore applied classification and predictive models for detecting attacks. On testing with the NSL-KDD data set, Random Forest Classifier outperformed the other approaches in identifying whether the data traffic was a normal one or an intrusion. Yassin et al. (2013) removed the drawbacks by applying an integrated algorithm called KMC + NBC (K-means clustering and Naïve Bayes classifier). After evaluating the empirical results, it showed that KMC+NBC lowered the false alarm rate but significantly improved the accuracy and detection rate. Abubakar and Pranggono (2017) enunciated a flow-based anomaly detection to outperform the weaknesses of signature-based IDS. Almost in all types of attacks, the results showed improvement for detection in all kinds of SDN environments with an alarmingly higher accuracy rate. Sultana et al. (2019) carried out a study and analyzed the emerging field of Software-Defined Networking (SDN). Various machine learning and deep learning approaches have been discussed, detecting vulnerabilities and supervising networks for any anomalies. Kumar, Thakur, and Ayyagiri (2020) presented a comprehensive review based on ensembles in machine learning. They also portrayed the current challenges and focused on future research work on developing effective IDS. Dang (2019) carried out a comprehensive survey of machine learning for IDS. He presented two techniques to detect the attacks in a network. First, a tree-based ensemble learning and a new method for selecting training data for IDSs using a small subset of training data combined with some unfavorable classification algorithms. The results obtained were better with low overhead. Peddabachigari et al. (2007) introduced two-hybrid approaches by integrating support vector machines (SVM) and Decision trees (DT) for modeling an efficient IDS. It was a hierarchical hybrid intelligent system model (DT–SVM) and an ensemble approach combining the base classifiers. This yielded in higher accurate detection rate and minimized computational complexity. Spark-Chi-SVM model was depicted by Othman et al. (2018) for detecting proliferations. KDD'99 was used to train and test, and the results were that the Spark-Chi-SVM model had high performance, reduced the training time, and the model was efficient for the big data environment. A new hierarchy anomaly intrusion detection model using fuzzy c-means (FCM) based on genetic algorithm and SVM was proposed by Tang et al. (2016). Empirical results showed that the model effectively detected the vast majority of network attack types. Even problems of false alarm rate and detection rate were also solved in the anomaly intrusion detection model. Tahir et al. (2015) proposed a hybrid machine learning technique using K-means clustering and support vector machine classification. It had a reduced rate of false-negative alarm, false positive alarm rate, and significantly improved detection rate. However, Li et al. (2012) integrated an ant colony algorithm, clustering method, and support vector machine (SVM) to detect whether an attack was normal or not in the network. The accuracy rate was quite higher using this integrated method. Khan et al. (2007) depicted a new approach integrating SVM and DGSOT. Rocchio Bundling technique and random selection method were compared with the said technique. Results achieved showed significant improvement over the Rocchio Bundling technique.

3. Problem Statement

The anomaly detection problem can be formulated as follows. A continuous stream of data $x \in R^k$ constitutes a collection of measurements $\{x_t\}_{t=1}^{T}$ governed by a probability distribution P. Although measurements correspond to specific physical events in the event space S, the mapping $f\colon S \to R^k$ between them may not be known. We assume that S can be divided into two subspaces corresponding to normal and anomalous physical events. A general approach to the problem mentioned above of learning such a representation consists of constructing a Minimum Volume Set with probability mass $\beta \in (0,1)$ concerning distribution P for a volume measure ξ :

$$G_\beta^* = argmin\{\xi(G) \colon P(G) \geq \beta, G\ measurable\}.$$

Real multidimensional data exhibit distributions that are highly sparse. Therefore, it is often desirable to reduce raw data's dimensionality via some feature extraction mechanism $g\colon R^k \longrightarrow R^l\ l < k$.

4. Objective

The objective here is to design an anomaly detector -

to show high accuracy and low false-positive rate (FPR) in detecting unknown attacks

to adapt to a continuously changing network environment

to compute faster and generate notification without any delay

5. Proposed Work

A hybrid Naïve Bayes-SVM model (NB-SVM) has been considered here that has two layers for classification. Naïve Bayes works well in real-time and consumes less time in computation; it is chosen as the first classifier. Initially, the incoming data or the anomalous data were first fed through the NB classifier, and the output is generated. The intermediate result is then again fed as input data to SVM. After going through the computation process of SVM, the final output is achieved. The final output differentiates between a normal attack and an anomaly.

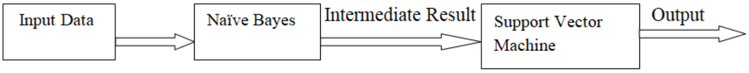


Fig 1: Block Diagram of the Hybrid Model

Naive Bayes classifiers are a group of classification techniques based on Bayes' Theorem. All of them share a common principle, i.e., every pair of features being processed for classification is independent of each other. It is not a single method. The fundamental rule of Naive Bayes is that each parameter plays an independent and equal role in the final result. Naive Bayes is probably one of the most simple, fast, and a popular learning approach. Naïve Bayes is considered here as layer 1 classifier due to the following reasons:

It is simple, easy, and less time-consuming.

In the case of multi-class prediction, performance is noteworthy.

Lesser amount of training data is required.

A Naive Bayes classifier offers promising results compared to other models like logistic regression techniques.

Both regression and classification challenges are solved by using a supervised machine learning approach called Support Vector Machine. Due to its popularity, it is best suited for classification problems. Each data item is plotted in an n-dimensional space (where n is the number of parameters), with each parameter's value representing a particular coordinate value. Classification is carried out by finding the hyper-plane that separates the two classes appreciably. Support Vectors simply coordinate individual observation.

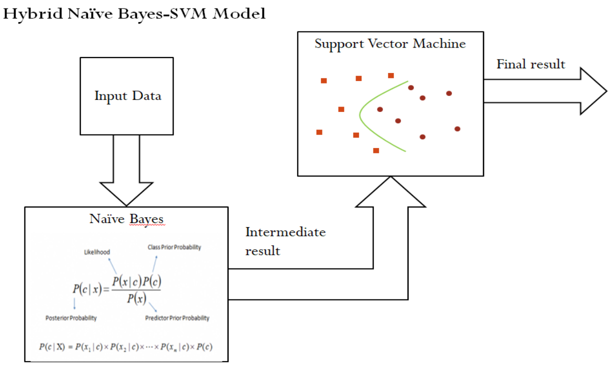

Hybrid Naïve Bayes-SVM Model
Input Data
Support Vector Machine
Final result
Naïve Bayes
Intermediate result
Likelihood
Class Prior Probability
$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability
$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$

Fig. 2 Two-Layer Classifiers

SVM has been used in layer 2 due to the following reasons:

can handle anomaly detection accurately outperforms unsupervised learning approach capability to deal with high dimensional data works well when integrated with various other techniques the decision boundary is determined by support vectors and extremely robust to outliers.

Although each classifier works better and yields promising results in individual cases, it may not significantly improve when the number of features and many training data are used. For this reason, the workability of the two-layer hybrid model is extended here with an ensemble approach.

All types of anomalies cannot be detected efficiently by a normal classifier model. Proper integration is required to increase the efficiency of the anomaly detector model. Ensemble approach is considered here to improve machine learning results by combining two single classifier models. It allows the generation of better predictive performance as compared to a single classifier or hybrid classifier. It also helps to overcome real-time challenges. It also solves the problem of detection rate and a high false-positive rate.

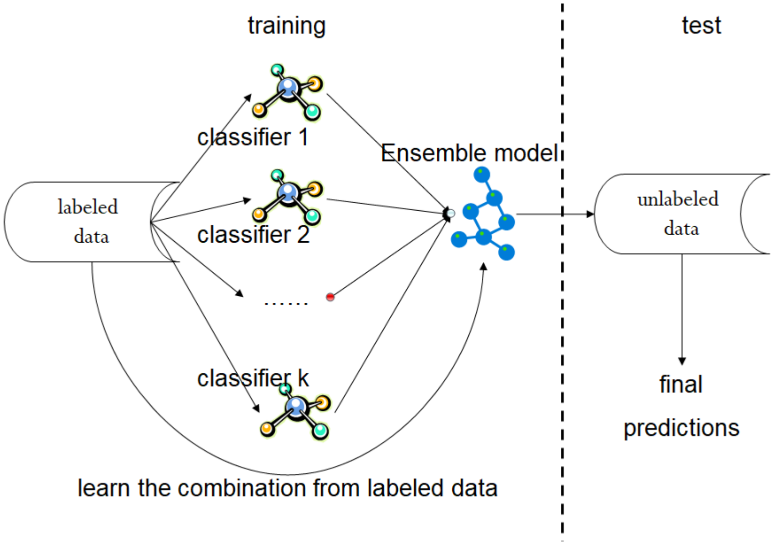


Fig. 3 Block Diagram 1 for Ensemble Approach

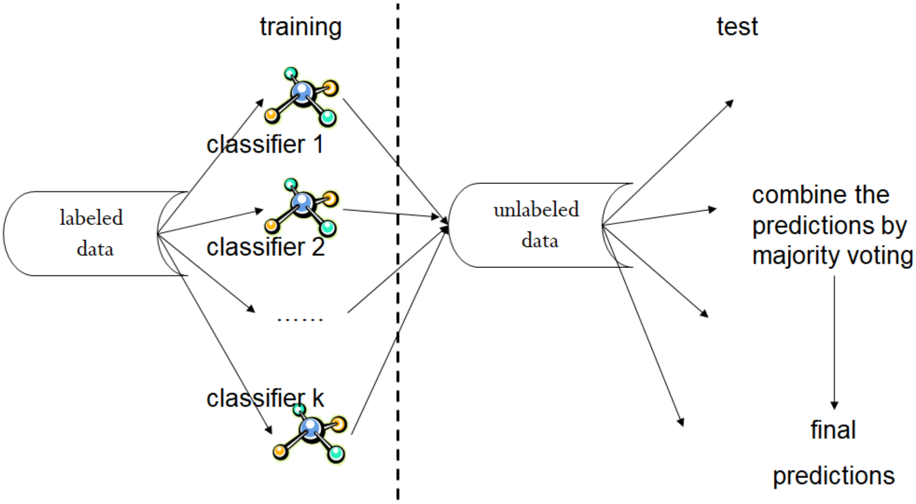


Fig. 4 Block Diagram 2 for Ensemble Approach

6. Discussions

Our preliminary results of applying machine learning techniques to network anomaly detection indicate their potential and highlight the areas where improvement is required. All parameters must be learned and autonomously set from arriving data to make the procedure portable to different applications and robust to diverse operating environments.

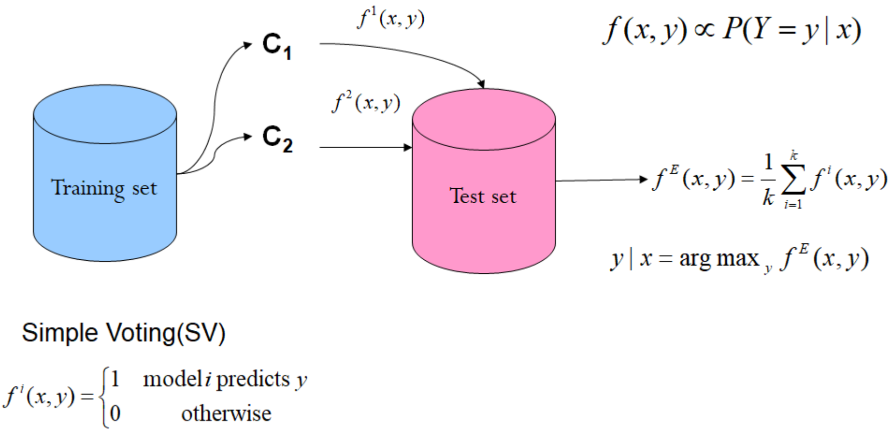
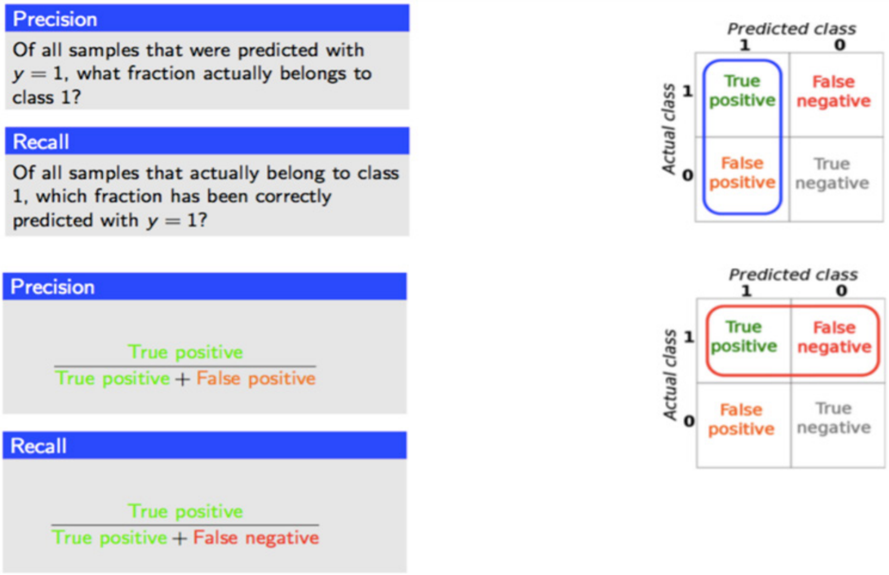

Training set
C₁
C₂
$f^1(x,y)$
$f^2(x,y)$
Test set
$f(x,y) \propto P(Y=y\mid x)$
$f^E(x,y)=\frac{1}{k}\sum_{i=1}^{k} f^i(x,y)$
$y\mid x=\arg\max_y f^E(x,y)$
Simple Voting(SV)
$f^i(x,y)=\begin{cases}1 & \text{model } i \text{ predicts } y\\ 0 & \text{otherwise}\end{cases}$

Fig. 5 Ensemble of C1 and C2 Classifier Model



Precision

Of all samples that were predicted with $y=1$, what fraction actually belongs to class 1?

Recall

Of all samples that actually belong to class 1, which fraction has been correctly predicted with $y=1$?

Precision

$$\frac{\text{True positive}}{\text{True positive}+\text{False positive}}$$

Recall

$$\frac{\text{True positive}}{\text{True positive}+\text{False negative}}$$

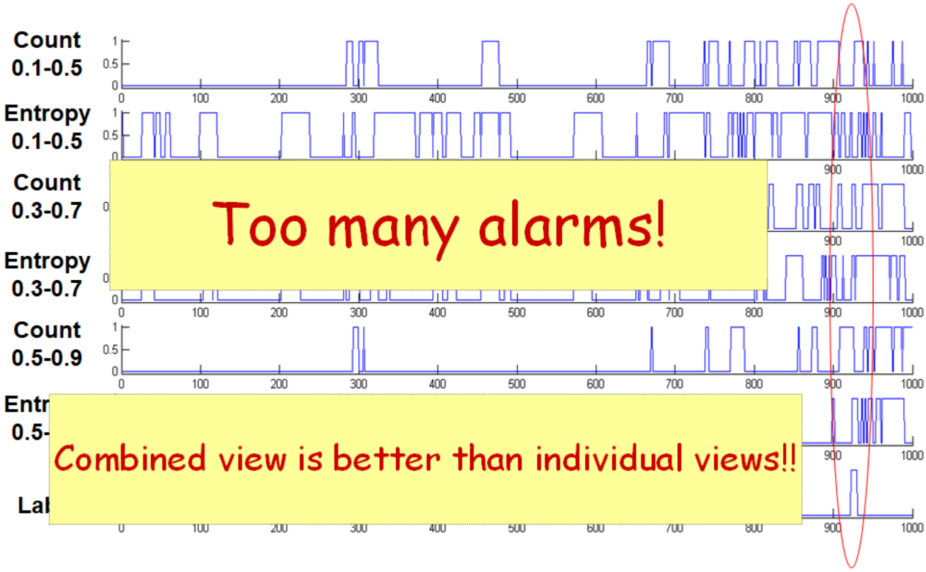
Actual class \ Predicted class	1	0
1	True positive	False negative
0	False positive	True negative

Actual class \ Predicted class	1	0
1	True positive	False negative
0	False positive	True negative

Fig. 6 Analyzing Parameters

We have initially tried KDD 1999 dataset to analyze the layer 1 classifier and layer 2 classifiers' performance. The following parameters are considered for detecting anomalous or normal events.

Figure 7 shows the entropy calculation for single classifiers and multiple classifiers. Anomaly detection using a single classifier generates too many alarms, which increases the entropy value of calculation. To make the procedure faster, two classifiers are integrated with an ensemble approach. Our current and future research focuses on rectifying these deficiencies in the approach and exploring other promising learning-based alternatives to the network anomaly detection challenge.



Too many alarms!

Combined view is better than individual views!!

Fig. 7 Entropy Calculation for Individual Classifiers

7. Conclusion

We have initially tried to carry out the classification process with a single classifier in our presented work. It is noted that it works best in individual scenarios. It may not yield noteworthy results in accuracy and computation when many features are present in the dataset. In such a scenario, the ensemble approach performs comparatively better, especially in handling vast network traffic where there is a challenging task of multiple types of anomalies and at the same time unknown. Here, the computation time is also less as Naïve Bayes works fast in real-time. Integrating multiple base classifiers, the ensemble approach provides promising detection results by compensating for their hazards. The results of this proposed work will enhance the performance of detecting anomalies with a low time frame. This method can be analyzed with real-time dataset to solve problems in a network environment as a future scope.

References

Abubakar, A., & Pranggono, B. (2017, September). Machine learning based intrusion detection system for software defined networks. In *2017 7th International Conference on Emerging Security Technologies (EST)* (pp. 138-143). IEEE.

Almseidin, M., Alzubi, M., Kovacs, S., & Alkasassbeh, M. (2017, September). Evaluation of machine learning algorithms for intrusion detection system. In *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)* (pp. 000277-000282). IEEE.

Belavagi, M. C., & Muniyal, B. (2016). Performance evaluation of supervised machine learning algorithms for intrusion detection. *Procedia Computer Science, 89*(2016), 117-123.

Dang, Q. V. (2019, November). Studying machine learning techniques for intrusion detection systems. In *International Conference on Future Data and Security Engineering* (pp. 411-426). Springer, Cham.

Esmaily, J., Moradinezhad, R., & Ghasemi, J. (2015, May). Intrusion detection system based on multi-layer perceptron neural networks and decision tree. In *2015 7th Conference on Information and Knowledge Technology (IKT)* (pp. 1-5). IEEE.

Haq, N. F., Onik, A. R., Hridoy, M. A. K., Rafni, M., Shah, F. M., & Farid, D. M. (2015). *Application of machine learning approaches in intrusion detection system: a survey. IJARAI-International Journal of Advanced Research in Artificial Intelligence, 4*(3), 9-18.

Jabez, J., & Muthukumar, B. (2015). Intrusion detection system (IDS): *anomaly detection using outlier detection approach. Procedia Computer Science, 48,* 338-346.

Khan, L., Awad, M., & Thuraisingham, B. (2007). *A new intrusion detection system using support vector machines and hierarchical clustering. The VLDB Journal, 16*(4), 507-521.

Kumar, G., Thakur, K., & Ayyagari, M. R. (2020). MLEsIDSs: machine learning-based ensembles for intrusion detection systems – a review. *The Journal of Supercomputing,* 1-34.

Leung, K., & Leckie, C. (2005, January). Unsupervised anomaly detection in network intrusion detection using clusters. In *Proceedings of the Twenty-eighth Australasian Conference on Computer Science-Volume 38* (pp. 333-342).

Li, Y., Xia, J., Zhang, S., Yan, J., Ai, X., & Dai, K. (2012). *An efficient intrusion detection system based on support vector machines and gradually feature removal method. Expert Systems with Applications, 39*(1), 424-430.

Mohamad Tahir, H., Hasan, W., et al. (2015). Hybrid machine learning technique for intrusion detection system. In *Proceedings of the 5th International Conference on Computing and Informatics* (pp. 464-472).

Omar, S., Ngadi, A., & Jebur, H. H. (2013). Machine learning techniques for anomaly detection: an overview. *International Journal of Computer Applications, 79*(2), 33-41.

Othman, S. M., Ba-Alwi, F. M., Alsohybe, N. T., & Al-Hashida, A. Y. (2018). Intrusion detection model using machine learning algorithm on Big Data environment. *Journal of Big Data, 5*(1), 34.

Peddabachigari, S., Abraham, A., Grosan, C., & Thomas, J. (2007). Modeling intrusion detection system using hybrid intelligent systems. *Journal of Network and Computer Applications, 30*(1), 114-132.

Ren, J., Guo, J., Qian, W., Yuan, H., Hao, X., & Jingjing, H. (2019). Building an effective intrusion detection system by using hybrid data optimization based on machine learning algorithms. *Security and Communication Networks,* 1-11.

Sabhnani, M., & Serpen, G. (2003, June). *Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context. In MLMTA* (pp. 209-215).

Sen, J. (2010, July). An intrusion detection architecture for clustered wireless ad hoc networks. In *2010 2nd International Conference on Computational Intelligence, Communication Systems and Networks* (pp. 202-207). IEEE.

Sinclair, C., Pierce, L., & Matzner, S. (1999, December). An application of machine learning to network intrusion detection. *In Proceedings 15th Annual Computer Security Applications Conference (ACSAC'99)* (pp. 371-377). IEEE.

Sultana, N., Chilamkurti, N., Peng, W., & Alhadad, R. (2019). Survey on SDN based network intrusion detection system using machine learning approaches. *Peer-to-Peer Networking and Applications, 12*(2), 493-501.

Tang, C., Xiang, Y., Wang, Y., Qian, J., & Qiang, B. (2016). *Detection and classification of anomaly intrusion using hierarchy clustering and SVM. Security and Communication Networks, 9*(16), 3401-3411.

Tsai, C. F., Hsu, Y. F., Lin, C. Y., & Lin, W. Y. (2009). Intrusion detection by machine learning: A review. *Expert Systems with Applications, 36*(10), 11994-12000.

Vinchurkar, D. P., & Reshamwala, A. (2012). A Review of Intrusion Detection System Using Neural Network and Machine Learning. *International Journal of Engineering Science and Innovative Technology, 1*(2), 54-63.

Wahba, Y., ElSalamouny, E., & ElTaweel, G. (2015). Improving the performance of multi-class intrusion detection systems using feature reduction. arXiv preprint arXiv:1507.06692.

Wang, G., Hao, J., Ma, J., & Huang, L. (2010). A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. *Expert Systems with Applications, 37*(9), 6225-6232.

Yassin, W., Udzir, N. I., Muda, Z., & Sulaiman, M. N. (2013, August). Anomaly-based intrusion detection through k-means clustering and naives bayes classification. In *Proc. 4th Int. Conf. Comput. Informatics, ICOCI* (pp. 298-303).

Zhong, S., Khoshgoftaar, T. M., & Seliya, N. (2007). Clustering-based network intrusion detection. *International Journal of Reliability, Quality and Safety Engineering, 14*(02), 169-187.

Submitted: 30.07.2020
Accepted: 23.11.2020