A Systematic Review on Anomaly Based Intrusion Detection System

R.Ganeshan¹, Chandra sekar kolli², Ch. Mohan kumar³ and T.Daniya⁴

¹Associate Professor, Dept of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh.
²Assistant Professor, Dept of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh.
³Assistant Professor, Dept of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh.
⁴Assistant Professor, Dept of IT, GMR Institute of Technology, Rajam, Andhra Pradesh, India

E-mail: daniya.t@gmrit.edu.in

Abstract. The Intrusion Detection System can monitor and investigate the harmful actions on a network and take necessary action to halt the external data. If the external request arrives, the IDS technique notifies the user about the intruders and thereby, allows the user to disconnect the IP network connection and reset the network connection from any interrupt. Hence, various anomaly based IDSs used for providing cyber security are reviewed in this paper. Accordingly, various research papers based on anomaly detection are reviewed by providing a classification based on several aspects, and are analyzed regarding various factors, such as datasets utilized, accuracy obtained, evaluation metrics, and implementation tools. From the analysis, it is observed that majority of the papers reviewed have employed MATLAB as the implementation tool.

Keywords: IDS, IP Network, Anomaly Detection, Cyber security

1. Introduction

IDS can be classified into five, such as host based IDS, network based IDS, Stack based IDS, signature based IDS, and anomaly based IDS. Figure 1 depicts the hierarchy showing different classifications of cyber security. The tree based IDSs include various techniques, like J48 and C4.5 Decision Tree algorithm, to address the classification problems that are common in the anomaly based IDS. Developed partial decision tree algorithm[1] through an
efficient Host based IDS – Correlation method. The attacks, such as DoS, U2R, R2L, and probing attacks, were handled by this method. The features were selected by the correlation feature selection method, and the partial decision tree had classified the packets of abnormal and normal behaviors. These combined algorithms were executed and tested using the feature selection method. In real time, the method could provide effective performance regarding detection time and accuracy. However, it fails to detect the unknown attacks.

Employed a decision tree[2], named as J48, which was constructed on real time traffic data to label every example like without intruders, known intruders and unknown intruders. These could be combined with the network intrusion detection to receive the higher accuracy rate in the J48 decision tree. [3] have developed two algorithms, named C4.5 Decision tree algorithm and C4.5 Decision tree with Pruning, for selecting the features. To automatically create the discrete values, the feature selection process was performed based on machine learning process. In particular, the features were classified to generate the rules. The limitation of this method is that it is not suitable for online packets. Described the security attacks [4] for IDS to detect the indifferences of network data. The performance of IDS was enhanced by selecting the features that were elaborated by the network data size. The system utilized the J48 classification algorithm, where the performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), kappa statistics measures, and root relative squared error. However, the benchmark of IDS needs further improvement.

![Figure 1: Cyber security through IDS](image)
2. REVIEW OF LITERATURE

2.1 Tree based IDSs

To utilize[5] the probabilities properties in an information system for the activity data that was used for discovering the intrusions. For intrusion detection, the probabilistic methods have included many techniques, such as chi-square multivariate test, decision tree, and Markov chain, Hotelling’s T2 test. In intrusion detection, the experiment was carried out using a sequence of multiple audit events in a frequency property, as a single audit event was insufficient for intrusion detection at a given period. Finally, in the frequency property, the intrusion detection technique produced a feasible structure with low computational energy and good performance level.

2.1.1 Advantage of Tree based method

- The tree based IDSs methods are easy in classifying the original and affected packets.
- The computational energy is reduced, while the performance level is maximized.
- The tree based methods have the ability to protect the data from the attacks that are not familiar and are used to identify the indifferences in network data.

2.2 Adaboost Algorithm based IDSs

The Adaboost algorithm, which is also a type of machine learning process, is used to enhance the benchmark. To develop online Adaboost-based intrusion detection algorithms [6], for classifying the weaker process by employing the online Gaussian mixture models and decision stumps. Here, the Particle Swarm Optimization (PSO) and local SVM-based algorithms were grouped to convert the detection models into global, since the IDS itself could not divide the features in each node. To designed IDS with less computational complexity based on the behavior of the network to maximize the rules, which were designed and learned from the classification. Hence, the classification rate could be increased successfully. Even though the rules generated had attained higher false alarm rate, the detection rate parameters require further improvement due to the complexity of the method. To developed an Adaboost-based algorithm for intrusion detection, where weak classifiers were utilized by the decision stumps. Therefore, the categorical and continuous features could be combined to produce the decision rules. The authors have employed the simple over fitting handling method to enhance the outcomes of the learning model. The major drawback of the method is the difficulty in the online updating model due to the existence of data streaming.

2.2.1 Advantages of Adaboost based IDSs

- The privacy of the data is protected, as the framework does not allow the actual data in the network.
- Higher classification rate can be achieved.

2.3 Probabilistic based IDSs

The Probabilistic based approach, such as Bayesian techniques, Conditional Random Fields and Layered Approach, Constant False-Alarm Rate (CFAR) detectors, and Hidden Markov Models (HMMs), used in anomaly based IDSs for the classification of the computational problem and the performance level is discussed. Developing [9] a predictive intrusion
detection model based on the Bayesian techniques and decision tree. The method could provide the best granularity level and positive rate for the IDS training model. To classify the problems occurred in the IDS, the model creation strategy had utilized the high priority data instances. Hence, the decision tree algorithm was trained effectively in every class for IDS. To [7] designed a Bayesian probability for IDS. To detect the possible intrusions, the system had utilized the Naive Bayes classifier. A superior detection rate could be attained during the intrusion detection by the Bayesian classifier. Thus, the system could detect the record type, employing the parallel Bayesian filters to enhance the accuracy of the IDS system.

Developed Conditional Random Fields and Layered Approach to address the efficiency and accuracy problems. The Conditional Random Fields could attain better detection accuracy, while the layered approach could offer good efficiency. The layered approach was utilized to minimize the estimation rate and to detect the invalid data. The limitation of the approach is that it cannot perform pipelining to execute the layers in the multi-core processors. To present a multiple network intrusion detection method depended on CFAR detectors and the data fusion techniques. Gaussian distribution was incorporated in the network traffic model to accept the defects that were predicted by the AR prediction linearity. This method could employ multiple CFAR detection techniques to solve the intrusion detection problem.

Developed IDS using HMM based on normal behaviors, and the most important differences were classified to identify the intrusions. For attaining adaptable and flexible requirements, the system employed the combination of two soft computing techniques, fuzzy logic and NN. To efficiently model the HMM as per the requirement, Self-Organizing map (SOM) was utilized that could minimize the optimal measures in audit data. In real time environment, the system could evaluate the intrusion detection with small scale trustable data.

2.3.1 Advantage of Probabilistic based IDS methods

- The approach can extract the best optimal solutions from the classifiers, providing the best accuracy level.
- The system is applicable to handle all the data, including the data with noisy information, and it is more trustable.

2.4 SVM based IDSs

The SVM based classification method is also a type of supervised learning process used to classify and analyze the selection or reduction set of the features in the network intrusion detection problems in an effective manner. Accordingly, various techniques, such as Ramp Loss K-Support Vector Classification-Regression (Ramp-KSVCNR), WOAR-SVM, and Apache storm, used for intrusion detection are explained. To developed a method, named Ramp-KSVCNR, to detect the problems due to multi-class intrusion. Besides, an alternative method, known as Alternating Direction Method of Multipliers (ADMM), which was applicable in large scale setting, was designed. Although the Ramp-KSVCNR method can minimize the trained time, it is more trustable to the third party. To [8] developed WOAR-SVM approach using a set of meta-heuristically generated weights to represent the relationship among the decision rules. Here, the input data was transformed into the decision rules, and the classification accuracy was maximized for the IDS. Accordingly, the multiclass problem was disintegrated into the single class problem to improve the
classification rate. The major drawback is that the SVM multiclass classification method cannot analyze the reduction set for the feature.

The author [9] have presented a network intrusion system and Apache storm framework, for analyzing high data loading. To revert and classify the high dimension data, the SVM utilized the statistical learning model. The pre-processed data was transformed over the nonlinear kernel with the help of SVM. The main impact of this method is that it requires a hybrid system to improve the efficiency in the intrusion detection task. We have developed an SVM based IDS employed with Principal Component Analysis (PCA) and without-PCA. The approach had classified the network traffic data in the IDS as original and misuse intrusions. We have presented a method, called weighted extreme learning machine employed model classification intrusion detection. To protect the data at the risk, the computer network system has utilized the NN. Moreover, the authors have used Weight adjustment by Weighted Extreme Learning Machine (Weighted ELM) technique, which was developed adopting the benefits of general ELM, to handle the imbalance data.

2.4.1 Advantages of SVM based IDSs

- These methods can provide better accuracy in detecting the intrusions in the network.
- SVM based IDSs can support the critical situations in the network to increase the system surveillances security effectively on time.

2.5 Neural Network based IDSs

The neural network can model the system with the interpretation of artificial intelligence through the Neuro - fuzzy classification method, DNN algorithm and Artificial Neural Network (ANNS) method, to find out the detection efficiency during the intrusion detection, as described. The authors [10] have developed an Artificial Intelligence IDS using Deep Neural Network (DNN). The input data was preprocessed, normalized, and transformed into the DNN model. The DNN algorithm was applied in IDS to refine the pre-processed data from the network. Finally, the DNN model had utilized various metrics, such as the quality, the rate for both detection and false alarm to find the detection efficiency for intrusion detection. The major defect was that the method could not categorize a single traffic data effectively. The Author have designed a hybrid structure with the combination of fuzzy logic and NNs, providing the efficient detection in a computer network. The inputs were pre-processed for the machine understandable language without any interruption and then, processed through the SOM block development. The hybrid structure has provided the best optimal features for the training model to distinct the well-known candidates.

To developed an intrusion detection framework using ANN method. Besides, the data was classified based on random weights using the NN and Radial Basis Function (RBF). Fisher feature selection algorithm was utilized to pre-process the data to eliminate the redundant data. The intrusion detection method developed had shown improved performance and so, it was accepted for the network data traffic in real-time applications. The authors had developed the Neuro-fuzzy model to minimize the false alarm rate due to its high efficiency and accuracy rate. The linguistic rules were interpreted in the Mamdani fuzzy inference system, to provide the best security level for the IDS selection process. Therefore, the misbehavior activities were minimized in the classification task.
2.5.1 Advantages of NN-based IDSs

- The intrusion detection methods developed using NN can improve the performance of IDS. Hence, it is accepted for the network data traffic in real-time applications.
- These methods minimize the misbehavior activities in the classification tasks for intrusion detection.

2.6 Unsupervised classification based IDSs

Different unsupervised techniques, such as K-Means clustering algorithm, unsupervised feature selection algorithm, and Multiple Adaptive Reduced Kernel Extreme Learning Machine (MARKELM), which are used to find the actual and the isolated packets in network intrusion, are described in the research papers discussed. The authors [11] have utilized K-Means clustering algorithm in the Distributed IDS. Accordingly, the data was analyzed and classified in the large scale network. The method was suitable for handling and monitoring the huge data in the network. The authors [12] have developed a method for intrusion detection in the network to differentiate the abnormal data packets from the original packets. The data mining approaches were combined with an operative approach to minimize the false alarm rate and to maximize the efficiency level. The authors [13,14] had designed an algorithm, named unsupervised feature selection algorithm, using the Laplacian score method.

The algorithm utilized the Extended Laplacian (EL) score based on the k-nearest neighbor graph and a Redundancy Penalization (RP) function. The Laplacian score was used to verify the effectiveness of the IDS. Hence, the locality power was preserved in each feature that was handled by the k-nearest neighbor graph and thereby, it would select the feature to reduce the redundancies in the data. The authors have developed a framework to enhance the data in network intrusion detection. Multiple learners were merged using Extreme Learning Machine (ELM) core algorithm, which was employed in the machine learning platforms to implement and to produce the best learning process in the SVM. Hence, the learning model can classify the network traffic, as original and affected packets, with low false alarm rate.

2.6.1 Advantages of Unsupervised classification based IDSs

- Unsupervised classification based IDSs are suitable for handling and monitoring the large scale data in the network.
- In this method, the false positive rate is minimized and an effort is made to enhance the efficiency level in IDS.

3. Conclusion

The research gaps identified in each approach used for detecting intrusions in the network is presented so that it will be useful for further research on the Anomaly based IDS. Moreover, an analysis is done based on the tools used for the implementation, evaluation metrics, accuracy ranges, and datasets employed, in each paper considered for the review. Based on the survey, it can be suggested that the majority of the research works had been implemented using MATLAB and the KDD CUP 99 dataset for the detection of the intrusions algorithm suggest the path to be taken in future works.
References

[1] K. Anand, J. Kumar, and K. Anand, “Anomaly detection in online social network: A survey,” 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Mar. 2017.

[2] V. A. Kotkar, “Scalable Anomaly Detection Framework in Video Surveillance Using Keyframe Extraction and Machine Learning Algorithms,” Journal of Advanced Research in Dynamical and Control Systems, vol. 12, no. 7, pp. 395–408, Jul. 2020.

[3] M. Nazam, M. Arshad, and A. Hussain, “Generalized Contractions on Dualistic Partial Metric Spaces,” Aug. 2016.

[4] N. M. Yasin, N. Balaji, G. Sambasivam, M. S. S. Basha, and P. Sujatha, “ADSMS: Anomaly Detection Scheme for Mitigating Sink Hole Attack in Wireless Sensor Network,” 2017 International Conference on Technical Advancements in Computers and Communications (ICTACC), Apr. 2017.

[5] V. P. K. Anne And D. R. Rao, “Standards and Analysis of Intrusion Detection based System: A Comparative Study,” Ponte International Scientific Researchs Journal, vol. 73, no. 2, 2017.

[6] T. V. Krishna Chowdary and K. V. V. Satyanarayana, “Data Gathering in Wireless Sensor and Actor Networks,” Indian Journal of Science and Technology, vol. 9, no. 31, Aug. 2016.

[7] K. V. S. N. R. Rao, S. K. Battula, and T. L. S. R. Krishna, “A smart heuristic scanner for an intrusion detection system using two-stage machine learning techniques,” International Journal of Advanced Intelligence Paradigms, vol. 9, no. 5/6, p. 519, 2017.

[8] U. Lakshmi Sowmya, M. Sai Kumar reddy, K. Madhu Babu, and D. K.V.D.Kiran, “Prevention of Spoofing Offensive in Wireless Sensor Networks,” International Journal of Engineering & Technology, vol. 7, no. 2.7, p. 770, Mar. 2018.

[9] R. Shaik and S. Shakeel Ahamad, “Security Attacks and Challenges of Wireless Sensor Network’s -a Review,” International Journal of Engineering & Technology, vol. 7, no. 2.32, p. 136, May 2018.

[10] R. Rajakumar, J. Amudhavel, P. Dhavachelvan, and T. Vengattaraman, “GWO-LPWSN: Grey Wolf Optimization Algorithm for Node Localization Problem in Wireless Sensor Networks,” Journal of Computer Networks and Communications, vol. 2017, pp. 1–10, 2017.

[11] B. S. Kumar, S. Karthi, K. Karthika, and R. Cristin, “A Systematic Study of Image Forgery Detection,” Journal of Computational and Theoretical Nanoscience, vol. 15, no. 8, pp. 2560–2564, Aug. 2018.

[12] B. S. Kumar, R. Cristin, K. Karthick, and T. Daniya, “Study of Shadow and Reflection based Image Forgery Detection,” 2019 International Conference on Computer Communication and Informatics (ICCCI), Jan. 2019.

[13] [24] C. C. Vignesh and S. Karthik, “Predicting the position of adjacent nodes with QoS in mobile ad hoc networks,” Multimedia Tools and Applications, vol. 79, no. 13–14, pp. 8445–8457, Apr. 2018

[14] B. Sathosh Kumar, M. P. Geetha, G. Padmapriya, and M. Premkumar, “AN APPROACH FOR IMPROVING THE LABELLING IN A TEXT CORPORA USING SENTIMENT ANALYSIS,” Advances in Mathematics: Scientific Journal, vol. 9, no. 10, pp. 8165–8174, Sep. 2020.