International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Intrusion Detection System using Soft Computing and Machine Learning Approach

Abhishek Vaidya[1], Vikrant Karawande[2], Kunal Gaikwad[3], Parshwa Shah[4], Anand Dhawale[5]

[1]*Student, Dept. of Computer Engineering, M.E.S. College of Engineering, Pune, India*
[2]*Student, Dept. of Computer Engineering, M.E.S. College of Engineering, Pune, India*
[3]*Student, Dept. of Computer Engineering, M.E.S. College of Engineering, Pune, India*
[4]*Student, Dept. of Computer Engineering, M.E.S. College of Engineering, Pune, India*
[5]*Asst. Professor, Dept. of Computer Engineering, M.E.S. College of Engineering, Pune, India*

Abstract - *Security is most important thing in software security and network security, basically any external entity accesses the resources without any authentication for authorization can be defined as a Intruder. Intruder can be internal for external which is generated from intentionally or generate automatically by any software. Many existing systems have already introduced intrusion detection systems (IDS) for network as well as host respectively. KDDCUP99 and NSLKDD data sets already proposed by the organization in 1999. Using multiple supervised learning algorithms system generates various signatures and policies to prevent the such anomalies into the vulnerable environment. In this paper we propose investigation of intrusion detection as well as prevention from different network attacks. Their existing data sets having a limitation to detect heterogeneous kind of attacks and we boost with different network attacks using various network data set. The propose system carried out data preprocessing, data normalization, feature extraction and feature selection before generate the training module. Once feature extraction has done it applies any supervise classifier for a training module. Similar process has executed in testing phase according to classification algorithm, and finally evaluates the classification accuracy for all attacks respectively. The proposed system has evaluated in Weka 3.7 open source environment with multiple supervise and unsupervised algorithms.*

Key Words: Intrustion Detection system, Soft computing, NIDS, HIDS, Fuzzy Rules.

1. INTRODUCTION

Various kind of attacks every day is faced by industries. And the best solution for it to use the Intrusion detection system (IDS). The computer networks security has been in the attention of investigation for years. The organization has come to realize that information & network security technology has become very important in protecting its information. Any successful attempt or unsuccessful attempt to compromise the integrity, confidentiality, and an accessibility of any information resource or the data itself is considered an intrusion or a security attack. As these are vulnerable to attacks, the broad use of computer networks and the enhance in business which are web based has made security of the network and host an significant issue. Attacks can be active or passive. Passive attacks only reads the data which is confidential and active attacks fabricates or modifies the data [7]. As it is not possible to keep away from these vulnerabilities and plan a totally secure system. Intrusion detection has turn into a major challenge. The main purpose of Intrusion detection system is to classify the attack and in a few cases examine it. Various approaches or techniques has been developed. But with the development of latest attacks more robust systems require to be designed. The contribution which we have made during the research project is as under:

- An effort has been made to provide a detailed better classification scheme, by which we will get an idea about the theoretical concept and working of various intrusion detection systems.
- We have provided a better method or model, which gave us understanding and proper guidance for selecting the appropriate intrusion detection system.
- An effort has been made to provide a detailed performance analysis of different intrusion detection system.
- An effort has been made to provide enhanced methods or new methods to mitigate the new type of attacks. By combining two different techniques to form hybrid technique.
- An effort has been made to implement the intrusion detection systems with the use of machine learning algorithms, which will learn the patterns and apply the same to test for intrusion detection.

2. LITERATURE SURVEY

Most of the researchers concentrate on genetic algorithm for creating the rules. For network intrusion detection, there are many proposed algorithm using well known KDDCUP99 dataset and only few are using real time network data. Some author uses GA for deriving classification rules and for providing optimal solution. Some author uses fuzzy algorithm for defining fuzzy membership function. There are various research papers related to IDS which has definite level of impact in computer and network security. According to Saeid Soheily Khah [1] proposed a system, Intrusion detection(ID) in networks is discussed through hybrid un supervised and unsupervised mining process - a thorough case study on the

ISCX benchmark dataset. It proposes a hybrid intrusion detection (kM-RF) which in general outperforms an alternative technique in terms of false alarm rate, accuracy and the detection rate. ISCX(A benchmark intrusion detection dataset) is used to assess the effectiveness of the kM-RF, and a deep analysis is performed to study the impact of the significance of every characteristic or features defined in the step of pre-processing. It also focuses on a dedicated pre-processing procedure to convert the categorical features or attributes to numerical ones and to build more isolated classes from the raw data, Few new features or characteristics to consider payloads, distributed attacks and an IP scans and A combination of k-means and random forest classifier to detect intrusion more effectively. The effectiveness of the suggested hybrid approach (kMRF) is analyzed on a dynamic, scalable and labeled benchmark dataset called as ISCX, which is the most up to date dataset compared to the other commonly explored ones for data intrusion benchmarking. The outcome shows the benefits of the kM-RF, which outperforms the other state of the art methods through the high accuracy, high detection rate and low false alarm rate, overall. A rank test signed by Wilcoxon is used to determine that the proposed kM-RF detection approach is significantly superior than the other methods.

According to Parisa Alaei et. Al. [2] in this approach, a method is proposed to overcome this problem by performing online classification on datasets. In doing so, an incremental naive Bayesian classifier is employed. Furthermore, active learning enables solving the problem using a small set of labeled data points which are often very expensive to acquire. The proposed technique is consisting of two groups of actions i.e. offline and online. The former involves data preprocessing while the latter introduces the NADAL online method. The proposed method is compared to the incremental naïve Bayesian classifier using the NSL-KDD standard dataset. There are three advantages with the proposed method: (1) overcoming the streaming data challenge; (2) reducing the high cost associated with instance labeling; and (3) improved accuracy and Kappa compared to the incremental naive Bayesian approach. Thus, the method is well-suited to IDS applications.

The fuzzy logic- based system can detect a malicious or intrusion behavior of a particular network, since it is rule based and it contains an improved set of rules. They have used automated approach for creation of fuzzy rules, obtained from the definite rules using Artificial items. The evaluations and experiment of the proposed intrusion detection (ID) system are executed with the well known KDD Cup 99 dataset. The proposed system achieved superior precision in identifying normal and intrusive records is shown clearly in the experimental result. The training kddcup dataset contains normal records and four different types of master attacks. For creating rules, system uses 10 different features. In the testing phase, the testing dataset is given as an input to the proposed system for classification of network normal or intrusive behavior. The final rules or output are afterwards used for detecting accuracy of the system based on recall, definition, precision, F-measures for estimating rare class prediction. Given system only working on Training and testing dataset it can't work on real time benchmark dataset. Given system shows the very good detection rate for all attacks it will not work for new signatures or attack [3].

The use of intrusion detection systems in soft computing techniques like neuro fuzzy and neural networks is used to classify network behavior and identify what category of attack got generated. For the initial categorization of the network traffic initially, neuro fuzzy classifiers are used. A system Fuzzy inference is later used to determine whether the activity is normal or abnormal. An IDS system is responsible for reducing false alarm rates. Human knowledge is used to create their fuzzy rule by Fuzzy inference systems. Genetic Algorithm is used for classification of network traffic in conjunction with ANFIS for obtaining best optimal solution. Genetic algorithms use a set of genetic parameters such as initialization of population, crossover, mutation rate, fitness and selection on current population to reproduce new optimal solution. There is a Poor detection rate for probe, U2R and R2L.System can create only static rules; it cannot work on dynamically new generated rules [4].

The fuzzy logic- based system is used for detecting an intrusion or malicious behavior of a particular network. Automated strategy is used for generating fuzzy rules. For detecting intrusion, the intrusion detection system which is proposed are evaluated with the use of dataset KDD Cup 99. The higher precision has been achieved by the intrusion detection system which is proposed in identifying whether the records are normal or malicious. The first component of the proposed system is to categorize input data or information into several classes depending upon different attacks involved in the intrusion detection. Second the designed strategy for automatic creation of fuzzy algorithm rules to give efficient learning. In general, fuzzy logic which has created the fuzzy rules are given by the fuzzy system by analyzing behavior of intrusion. The procedure of fuzzy generation has given as follows.

- Mining of only length Artificial items.
- Rule generation
- Rule filtering
- Generating fuzzy rules

The old fitness function for Genetic Algorithm is used. The apriori algorithm is used for association rules for increasing the system time complexity and execution time [5].

Intrusion Detection (ID) with Genetic Algorithm (GA) and fuzzy Logic describes two different behavior of training intrusion detection (ID) system to recognize possible attacks in a network particularly or computer system. It will illustrate an approach by using fuzzy genetic algorithms and evaluate those records with rules obtained using a decision tree. Genetic algorithm uses genetic parameters like crossover, mutation, selection for obtaining optimal solution. The GA rules which are generated by Genetic algorithm are given as an input to Fuzzy logic for classification of master a Section 1: described the procedures that are used in determining the accurateness rate of IDS. Section 2: described the fuzzy genetic(FG) algorithm in IDS. Section 3: described the outcome of using a conventional genetic algorithm.

The detection rate for a proposed system is approximate 98.00%. There are numerous restrictions to the prevention-based approach for network and computer security. It is almost certainly not possible to construct a totally secure system. The prevention-based security viewpoint constrains the user's productivity and activity [6].

Intrusion Detection (ID) System using fuzzy genetic(FG) algorithm (FGA) is used to classify network attack. The proposed approach evaluates intrusion detection (ID) system into false positive alarm, detection rate and detection speed. Fuzzy genetic algorithm (FGA) can classify two activities i.e. malicious and normal behavior. For each generation, population size of 10 is considered. Mutation rate of 30% and single point crossover has applied. Online network dataset can be detected and evaluated by fuzzy genetic algorithm, within 2 to 3 seconds. Preprocessing requires 2 seconds and fraction of second is required for detection. With low false positive rate and high accuracy, fuzzy genetic algorithm can detect recent network activities using online dataset and KDDcup dataset. The rate of detection is over 97.5% .System cannot detect unknown attack or the attacks whose signature is not predefined. System cannot generate dynamic rules [7].

According to [8], a system to accurately detect potential attack has developed by using various techniques like decision free, Random forest and KNN. To overcome the limitation of the previous system that was not able to detect the IPV6 attacks, a new method are proposed. The developed system produce the impressive and efficient result in identifying IPV4-based attack keeping in mind the future scope. The effectiveness of various algorithm evaluated. Detection accuracy, precision, recall percentage were measured.

According to [9] has stated that clustering and KDD can be efficiently used to detect novel anomaly called NEC. An unsupervised anomaly is used to produce high detection rate and less false passive rate. It is an appropriate way to solve the difficulty and find the anomaly which does not need a labeled data set. The system is verified over NSL-KDD 2009 dataset. The preprocessing model transforms all features into the real number and normalized dataset at the end the evaluation component will compare predicate result an accurate result.

Concerning A Survey of Data Mining and Machine Learning for CSID [10], a survey of data mining and machine learning for cyber security instruction detection is performed to ensure cyber security. Packet-header and net flow packet header are used for the instruction detection system to be able to reach networks and kernel level data. The future scope that is kept in mind is that data mining and machine learning cannot ware without representative data and also it's very time-consuming. The complexity of different machine learning and data mining algorithm is discussed, The research paper provides a set of comparison criteria for machine learning/data mining methods Intrusion Detection System help discovered, determine and identify unauthorized used, duplication, alteration and destruction of the information system.

3. RESEARCH METHODOLOGY

The IDS has categorized into the two different sections like network based intrusion detection system (NIDS) and host based intrusion detection system (HIDS). Basically both systems deals with various machine learning as well as soft computing algorithms. In this propose implementation we utilized various network intrusion data set which taken from the per organizations. The 41 attributes already available in entire data set during the data preprocessing we validate each attribute value with desire ranges then normalize the data using attribute selection technique. The system deals with first 6 attribute which holds multi-value or categories values.

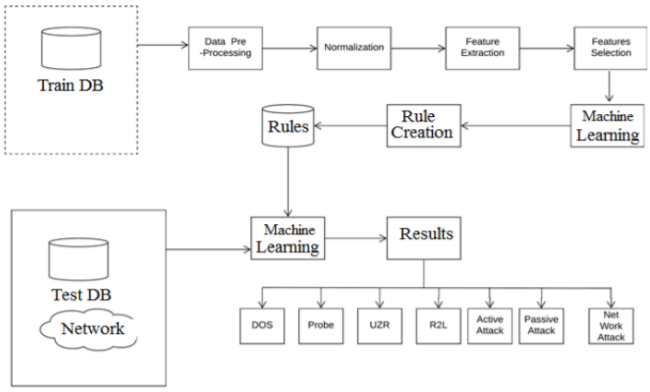

Fig -1: Proposed System Architecture

For classification multiple classification algorithms has exist in Weka tool environment. Some supervised learning, un-supervised learning as well as various reinforcement learning algorithms are already available in clustering section. The process food validation option also available after selection the respect to classifier. The three different parameter tuning options are available like used as a training data set, cross validation (we can choose like 5 fold, 10 fold etc) and finally provide both data set in separate file section. Once attribute selection pattern we can build the classifier. Once complete the entire execution system shows confusion matrix according to given input. Entire matrix would be show precision, recall, accuracy and f-measure score respectively.

3.1 Module 1: Input Dataset

The KDD cup 99 dataset contains several statistical analyses which affects to detect the accuracy of many IDS model. The KDD dataset has training and testing dataset. The total number of training dataset is approximately 4, 900, 000. It contains 41 features with labeled as normal or specific attack type. 300,000 instances contains in testing dataset with twenty-four training attack types and extra fourteen attack types in the test set only. NSL-KDD data set is a modified version of its precursor. It consists, important records of the KDD dataset. There are 4 attack classes of anomaly which are further classified as DoS, Probe, R2L and U2R. In that there are subtypes of each anomaly attack types.

3.2 Module 2: Data Pre-processing

Data pre-processing done by Weka tool. This is the offline method. Data pre-processing includes following three main tasks:

a. Converting non-numerical features of NSL-KDD dataset into numerical values.

b. At the end transferring attack types into numerical values.

c. Finally preparing proper dataset.

3.3 Module 3: Training and Clustering

Step 1: Upload training data for feature extraction.

Step 2: Apply discretization approach on dataset.

Step 3: Apply Normalization approach on dataset.

Step 4: Generate normalize and discretize dataset.

3.4 Module 4: Testing Dataset

Step 1: Upload Testing data for detection.

Step 2: Apply discretization approach on dataset.

Step 3: Apply Normalization approach on dataset.

Step 4: Generate normalize and discretize dataset.

Step5: Apply ensemble approach {NB, J48, ANN}

Step 6: classify all attacks.

Step 7: Show detection results.

4. DATASET DESCRIPTION

The inherent drawbacks in the KDD cup 99 dataset [4] has been revealed by various statistical analyses has affected the detection accuracy of many IDS modeled by researchers. It contains essential records of the complete KDD data set. There are a collection of downloadable files at the disposal for the researchers.

Table -1: Dataset Description

Id	Name	Description
1	KDDCUP99	41 Attributes with 23 sub classes for all 4 classes.
2	NSLKDD	41 Attributes with 38 sub classes for all 4 classes.
3	Botnet	12 attributes including class as normal and abnormal
4	ISCX	29 attributes including class as normal and abnormal
5	NUSW-NB15	It contains 49 attributes binary, 0 for normal and 1 for attack records
6	WSNtrace	12 attributes including class as normal and abnormal

5. RESULTS AND DISCUSSION

To evaluate the proposed system performance analysis on weka 3.7 simulation environment, we have used various data set for system testing which is already define in table 1. Each data set contains different features as well as different kind of attacks. Once the system has train according to specific data set, it generates training rule accordingly. The average accuracy for entire system with all data set is around 90%.

In our experimental setup we have done various experiments, the confusion Matrix has been calculated for each data set according to label assign by testing algorithm. The testing data set which is basically and label when we deals with the system testing. The classification accuracy should we generate according to two given threshold, the threshold value has set initially 0.70. The optimum threshold for this research it's around 0.60, which displays better accuracy than others. The proposed multiple machine learning algorithms provides the classification which is shown in below figure 1 to figure 3.

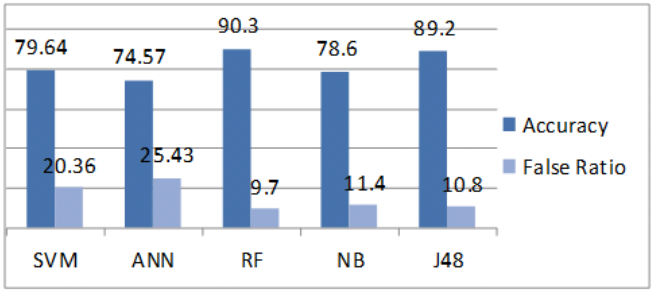

Chart -1: System classification accuracy when use as training dataset with KDDCUP99 dataset

According to above chart 1, we used same dataset for training as well testing in first experiment, all five classifiers has executes concurrently. The Random forest provides 90.30% highest accuracy for all attack classes while ANN provides 74.57% lower accuracy respectively.

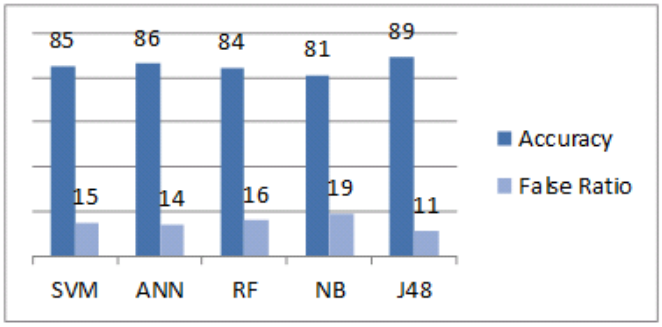

Chart -2: System classification accuracy with supplied test dataset with KDDCUP99 dataset

According to above chart 2, we used supplied train and test dataset for training as well testing in second experiment, all five classifiers has executes concurrently. The J48 provides 89% highest accuracy for all attack classes while NB provides 81% lower accuracy respectively.

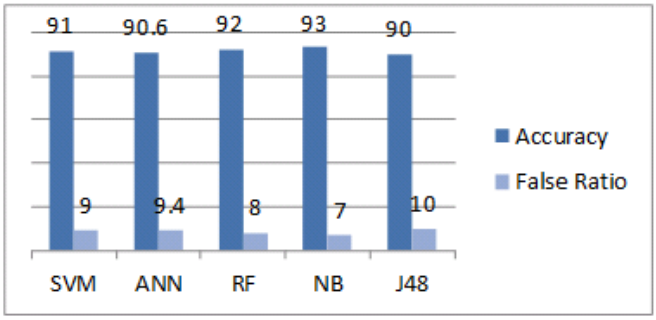

Chart -3: System classification accuracy 10 fold cross validation with KDDCUP99 dataset

According to above chart 3, we used various cross fold validation (10 fold cross validation) for training as well testing in third experiment, all five classifiers has executes concurrently. The NB provides 93% highest accuracy for all attack classes while J48 provides 90% lower accuracy respectively.

We did not compare our algorithm with other malware detection algorithms because our binary and malicious files did not match the format of the files required to run these algorithms. Additionally, it did not make sense to compare the accuracy between algorithms tested on different datasets. As a result, we compared our algorithm to other state of the art machine learning classifiers using the data obtained by extracting certain features from the original binary and malicious files. We tested the data with the support vector machine classifier [8], ANN classifier [9], and the RF classifier [10].

The accuracy produced by the other machine learning classifiers significantly varied across the individual feature sets. For the DLL feature sets, the other machine classifiers performed roughly around the same as our classifier. Our classifier performed better than the support vector machine, ANN, and ensemble classifiers for the byte sequences feature set. Finally, all three of the other machine learning classifiers performed much better than our classifier on the strings feature set. However, our combined classifier still yields the highest accuracy of overall accuracy.

6. CONCLUSION

The proposed implementation has shown classification accuracy for selective input data set with various soft computing as well as machine learning algorithms. With the various experiment analyses finally we conclude system takes different time complexity to execute the respective classification algorithm. System deals with different parameter tuning process during the train as well as test classifier. Multiple algorithms provide highest classification accuracy then some classical intrusion detection systems while some new feature selection strategies also improve the classification accuracy. The average accuracy for entire system Naive Bayes should be provides highest classification accuracy while SVM introduces lower accuracy than other classical algorithms. To evaluate the proposed system with multiple network intrusion data set for network as well as software security with deep learning algorithms will be the future direction for this system.

REFERENCES

[1] Sedjelmaci H, Senouci SM, Ansari N. A hierarchical detection and response system to enhance security against lethal cyber-attacks in UAV networks. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2018 Sep;48(9):1594-606.

[2] Alaei P, Noorbehbahani F. Incremental anomaly-based intrusion detection system using limited labeled data. In Web Research (ICWR), 2017 3th International Conference on 2017 Apr 19 (pp. 178-184). IEEE.

[3] R. Shanmugavadivu, "Network Intrusion Detection system using Fuzzy logic", ACM Digital Library, Volume 30 Issue 1, January 2007.

[4] Emma Ireland, "Intrusion Detection with Genetic Algorithms and Fuzzy Logic", UMM CSci Senior Seminar Conference, Morris, MN, December 2013.

[5] Rupesh B. Jadhav and Mr. Balasaheb B. Gite, "Real Time Intrusion Detection With Fuzzy, Genetic and Apriori Algorithm", International Journal of Advance Foundation and Research in Computer (IJAFRC), Vol 1, Nov 2014.

[6] S. N. Pawar, "Intrusion detection in computer network using FGA", IEEE journal on parallel and distribute systems, Vol.23, No.3, March 2012.

[7] P. Jongsuebsuk, N. Wattanapongsakorn and C. Charnsripinyo,"Real-Time Intrusion Detection with Fuzzy Genetic Algorithm", IEEE 2013.

[8] Mohammed Anbar, Rosni Abdulah, Izan H. Hasbullah, Yung- Wey Chong; Omar E. Elejla, "Comparative Performance Analysis of classification algorithm for Internal Intrusion Detection ", 2016 14th Annual Conference on Privacy Security and Trust (PCT), Dec 12-14,2016, Penang, Malaysia.

[9] Weiwei Chen, Fangang Kong, Feng Mei, GuiginYuan, Bo Li, "a novel unsupervised Anamoly detection Approach for Intrusion Detection System", 2017 IEEE 3rd International Conference on big data security on cloud, May 16-18,2017, Zhejiang, China.

[10] Anna L. Buczak, Erha n Guven, "A Survey of Data Mining and Machine Learning methods for cybersecurity intrusion detection", IEEE communication surveys and tutorials, vol. 18, Issue 2,2016.

[11] Soheily-Khah, Saeid, Pierre-François Marteau, and Nicolas Béchet. "Intrusion Detection in Network Systems Through Hybrid Supervised and Unsupervised Machine Learning Process: A Case Study on the ISCX Dataset." Data Intelligence and Security (ICDIS), 2018 1st International Conference on. IEEE, 2018.

[12] Mohammed Hasan Ali, Bahaa Abbas Dawood AL Mohammed1, Madya Alyani Binti Ismail, Mohamad Fadli Zolkipli, A new intrusion detection system based on Fast Learning Network and Particle swarm optimization.