ISSN 2278-3091

Volume 9, No.4, July – August 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse321942020.pdf

https://doi.org/10.30534/ijatcse/2020/321942020

Intrusion Detection System Various Supervised Learning Techniques

Rashmi Bhaskarrao Kale[1], Dr. Shafi Pathan[2]

[1]Research Scholar, India, rashmi.kale2705@gmail.com

[2]Professor, India, shafipathan@gmail.com

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 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 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 pre-processing, 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: Intrusion Detection system, Soft computing, Deep Learning, RNN, NIDS, HIDS

1. INTRODUCTION

Industries encounter unique kinds of attacks each day. And the perfect Intrusion Detection System (IDS) solution for it. Security computer networks has been at the prosecution's attention for years. The organization has come to realize that security technology in information & networks has become quite important in protecting their information. Every successful attempt or unsuccessful attempt to compromise the privacy, confidentiality and accessibility of any information resource or the data itself is considered an intrusion or a security assault. Because these are vulnerable to attacks, the widespread use of computer networks and the enhancement of web-based business has made network security and host a major issue. Can be aggressive or passive assaults. Passive attacks read only certain data that are confidential and active attacks that produce or alter data [7]. Since such flaws cannot be held away and a completely stable device can be planed. Detection of the intrusion has become a major challenge. Intrusion detection system's main objective is to identify the attack and analyze it in a few instances. We have developed various methods or techniques. Yet more reliable structures involve planning with the advent of the latest attacks.

The contribution we made during the research project is as follows:

- Efforts have been made to include a detailed better classification scheme which will give us an idea of the theoretical definition and the operation of various intrusion detection systems.
- We have given a better approach or model that gave us understanding and guidance to choose the best framework for intrusion detection.
- A detailed performance analysis of the different intrusion detection systems has been undertaken.
- An attempt has been made to include effective techniques or alternative methods to reduce the latest form of attacks. Combining two techniques to form a hybrid technique.

There has been an effort to incorporate detection mechanism using machine learning algorithms that will learn these same patterns as well as apply the same to intrusion identification testing.

2. LITERATURE SURVEY

The majority of researchers rely on the genetic algorithm to construct the rules. There are several proposed algorithms using well-known KDDCUP99 dataset for network intrusion detection and a few use real time network data. Some author uses GA to derive rules for classification and to have an optimal solution. Many writers use fuzzy algorithms to describe fuzzy membership function. There are many IDS-related research papers which have a definite degree of effect on computer and network security. According to a method proposed by Saeid Soheily Khah[1], intrusion detection(ID) in networks is discussed via hybrid both controlled and unmonitored mining phase-a detailed case study on the ISCX benchmark dataset. This proposes a detection of hybrid intrusion (kM-RF) which generally outperforms an alternative technique in terms of false alarm rate, accuracy and rate of detection. ISCX(A benchmark intrusion detection dataset) is used to assess the effectiveness of kM-RF and an in-depth analysis is carried out to test the impact of the significance of any characteristics or characteristics identified in the pre-processing stage. It also focuses on a dedicated pre-processing method to convert categorical features or attributes to numerical features and create more isolated classes from raw data, few new features or features to recognize payloads, distributed attacks and IP scans, and a combination of k-means and random forest classifier to more effectively detect intrusion. The efficacy of the suggested hybrid method (kMRF) is tested on a complex, scalable and labelled benchmark dataset called ISCX, the most up-to - date dataset compared to other widely studied data intrusion benchmarking datasets. The result shows the advantages of the kM-RF, which outperforms the other state-of-the-art methods through the overall high precision , high detection rate and low false alarm rate. A Wilcoxon-signed rank check is used to assess that the proposed kM-RF detection strategy is significantly superior to the other approaches.

According to Alaei et. of Parisa In this approach, Al.[2] proposes a method for dealing with this problem by performing online classification on datasets. By doing so, it uses an incrementally naive Bayesian classifier. Moreover, active learning enables the problem to be solved by means of a small set of labeled data points that are often very expensive to acquire. The proposed technique consists of two acting groups , i.e. offline and online. The former includes pre-processing of data, while the latter uses the online NADAL system. The approach suggested is contrasted with the incremental naive Bayesian classifier using the standard NSL-KDD dataset. The proposed method has three advantages : (1) overcoming the data streaming challenge; (2) reducing the high cost associated with instance labeling; and (3) improving accuracy and Kappa as compared to the incremental naive Bayesian approach. Therefore the method is suitable for IDS applications.

The fuzzy logic-based system can detect a specific network's malicious or interference actions because it is built on guidelines and provides an improved set of laws. We used an automated method for constructing fuzzy rules, achieved using Artificial objects from the definite rules. The evaluations and experiments of the proposed system of intrusion detection (ID) are carried out with the well-known dataset KDD Cup 99. The proposed method obtained superior precision in the detection of natural and intrusive records as clearly seen in the results of the experiment. The data set for the kddcup training includes normal records and four different types of master attacks. Program uses 10 different features for developing rules. The research dataset is provided in the testing process as an input to the proposed method for classifying normal or disruptive activity in the network. Afterwards, the final rules or performance are used to detect system accuracy based on recall, classification, precision, F-measures to estimate unusual class prediction. Since the program only operates on the Training and Testing dataset it can not operate on the benchmark data set in real time. Since device shows the very good rate of detection for all attacks, it won't work for new signatures or attacks[3].

Using intrusion detection systems in soft computing techniques such as neuro fuzzy and neural networks is used to classify the behavior of the network and identify which category of attack has been generated. Initially, neuro fuzzy classifiers are used for the initial categorisation of network traffic. A Fuzzy inference system is later used to determine if the activity is normal or anomalous. It is for an IDS system to reduce false alarm rates. Human knowledge is used by Fuzzy inference systems to create their dubious rule. The Genetic Algorithm is used in conjunction with ANFIS for the classification of network traffic to obtain the best optimum solution. Genetic algorithms use a set of genetic parameters for reproducing new optimal solution, such as population initialization, crossover, mutation rate, fitness and selection on current population. The probe has a poor detection rate, U2R and R2L.System can only create static rules; it can't work on dynamically new generated rules[4].

Fuzzy logic-based program is used to detect a specific network intrusion or malicious behavior. Automated strategy is used to generate flippant rules. To detect intrusion the proposed intrusion detection system is evaluated using the KDD Cup 99 dataset. The higher precision was obtained through the method of intrusion detection which is proposed to determine whether the records are regular or malicious. The first component of the proposed system is to categorize input data or information into several classes, depending on the various intrusion detection attacks. Second, the strategy designed to offer successful learning to the automatic development of fuzzy algorithm rules. In general, the fuzzy method provides fuzzy logic that has generated the fuzzy rules by evaluating intrusion behaviour. The Fuzzy Generation procedure has provided the following.

- Mining of Artificial Objects only in length.
- Building rules
- Filtering Law
- Generate flippant rules

For Genetic Algorithm the old fitness function is used. The apriori algorithm is used for rules of association to increase the complexity of machine time and execution time[5].

The Genetic Algorithm (GA) and Fuzzy Logic intrusion detection (ID) method defines two separate training intrusion detection (ID) behaviors to identify possible attacks in a particular network or computer device. Using fuzzy genetic algorithms it will illustrate an approach and evaluate those records with rules obtained using a decision tree. Genetic algorithms use genetic parameters such as fusion, mutation, selection to get the optimal solution. The GA rules generated by the Genetic algorithm are given as an input to Fuzzy's section 1 master classification logic: describing the procedures used to determine IDS' accuracy rate. Section 2: Fuzzy genetic(FG) algorithm described in IDS. Section 3: The results of using a conventional genetic algorithm are described.

The detection rate is approximately 98.00 percent for a proposed system. There are several limitations on the network and data security focused approach to prevention. Construction of a fully secure system is almost certainly not possible. The security viewpoint based on prevention constrains the productivity and activity of the user[6].

Systems using fuzzy genetic(FG) algorithm (FGA) intrusion detection (ID) are used to identify network attacks. The proposed approach assesses system for intrusion detection (ID) into false positive alarm, rate of detection and speed of detection. Fuzzy genetic algorithm (FGA) can distinguish two behaviors, i.e. natural and malicious behaviour. Population size is considered at 10 for each generation. A 30 per cent mutation rate and a single point crossover has been applied. Online network dataset can be detected and evaluated within 2 to 3 seconds, using a fuzzy genetic algorithm. Preprocessing takes 2 seconds, and detection requires fraction of a second. Using online dataset and KDDcup dataset, fuzzy genetic algorithm can detect recent network activities with low false positive rate and high precision. The detection rate exceeds 97.5 percent. The system can not detect unknown attacks or attacks that are not predefined by their signature. Program can't create complex rules[7].

According to [8], a system has been developed to accurately detect potential attacks using various techniques such as free decision, random forest and KNN. A new method is proposed to overcome the limitation of the previous system that could not detect IPV6 attacks. The developed system produces the impressive and efficient result of identifying an IPV4-based attack with the future scope in mind. Assessed the efficiency of different algorithms. Accuracy of detection, precision and percentage of recall were measured.

Clustering and KDD can be used efficiently to detect novel anomaly called NEC, according to [9]. To achieve high detection rate and less false passive rate, an unsupervised anomaly is used. It's an easy way to fix the issue and locate the anomaly that doesn't require a classified collection of data. The device is tested over the 2009 NSL-KDD dataset. The preprocessing model converts all functionality into the actual number and the structured data set will be compared to the predicate result at the end of the evaluation section..

With respect to a based on machine learning survey for CSID[10], cyber security instruction detection is conducted to ensure information security. For the application detection system, the post processing and the net flow packet header are used to attain the networks and processor level data. The potential area that is kept in mind is that through representative data , data analysis and machine learning could not ware and is therefore very time-consuming. The difficulty of various machine learning and data algorithms is addressed, The research article presents a set of comparative criteria for machine learning / data mining approaches Intrusion Detection System accurately identify, assess and recognize inappropriate use, replication, modification and destruction of the information system.

The multi class classification problem has solved using various machines learning algorithm in [11] [12]. Both system deals with large unstructured data and proceed using supervised learning approach. The achieved results demonstrated they got better classification accuracy than traditional machine learning algorithms.

3. PROPOSED SYSTEM DESIGN

3.1 : System Design

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 utilised 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.

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.

3.2 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

4. ALGORITHM DESIGN

4.1 Training algorithm for rule generation

Input: Training dataset TrainData[], Various activation functions[], Threshold Th
Output: Extracted Features Feature_set[] for completed trained module.

Step 1: Set input block of data d[], activation function, epoch size,
Step 2 : Features.pkl ← ExtractFeatures(d[])
Step 3 : Feature_set[] ← optimized(Features.pkl)
Step 4 : Return Feature_set[]

4.2 Algorthm for system testing

Input: Training dataset TestDBLits [], Train dataset TrainDBLits[] and Threshold Th.
Output: Resulset <class_name, Similarity_Weight> all set which weight is greater than Th.
Step 1: For each testing records as given below equation

$$testFeature(k) = \sum_{m=1}^{n} (.\ featureSet[A[i] \ldots\ldots A[n] \leftarrow TestDBLits)$$

Step 2 : Create feature vector from $testFeature(m)$ using below function.

$$\text{Extracted\_FeatureSet\_x } [t\ldots\ldots n] = \sum_{x=1}^{n}(t) \leftarrow testFeature\ (k)$$

Extracted_FeatureSet_x[t] holds the extracted feature of each instance for testing dataset.
Step 3: For each train instances as using below function

$$trainFeature(l) = \sum_{m=1}^{n} (.\ featureSet[A[i] \ldots\ldots A[n] \leftarrow TrainDBList)$$

Step 4 : Generate new feature vector from $trainFeature(m)$ using below function
.

$$\text{Extracted\_FeatureSet\_Y}[t\ldots\ldots n] = \sum_{x=1}^{n}(t) \leftarrow TrainFeature\ (l)$$

Extracted_FeatureSet_Y[t] holds the extracted feature of each instance for training dataset.
Step 5 : Now evaluate each test records with entire training dataset

$$weight = calcSim\ (FeatureSetx \,||\, \sum_{i=1}^{n} FeatureSety[y])$$

Step 6 : Return Weight

5. RESULTS AND DISCUSSIONS

To verify the proposed performance evaluation on the simulation model for weka 3.7, we used various data sets for system testing which are already specified in table 1. Through collection of data includes various features and also different kinds of attacks. Once the system has train according to specific data set, training rule is developed accordingly. The average accuracy of all data set for both the entire system 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 and figure 2.

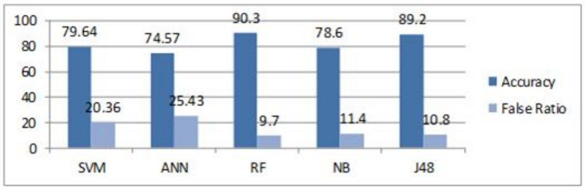

Figure 1: System classification accuracy when use as training dataset with KDDCUP99 dataset

According to above figure 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.

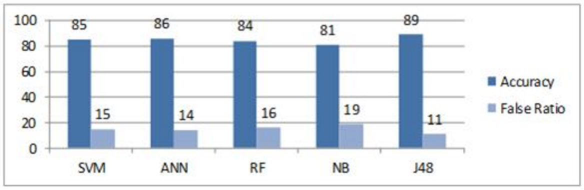

Figure 2: System classification accuracy with supplied test dataset with KDDCUP99 dataset

According to above figure 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

The above chapter describes experiment analysis and result analysis of system, in two section provided implementation execution and result calculation of system with proposed algorithm and comparison with existing algorithms. Data extraction form TFL and store into middleware Linode cloud system provides like Service Oriented Architecture (SOA) to system. Various machine learning base supervised algorithms evaluated in system and finally conclude the result section of entire research

6. CONCLUSION

The research's main objective is to build a Cloud data model for a multi - modal transportation process that integrates multiple modes with one network, enabling various modal compositions in traffic management. The optimization technique adopted for this research to achieve these objectives is mainly to provide a separate entity with each mode route and to distinguish each of these identities functionally. In various methods the distinction is fixed and variable for the various routes of the same mode. Getting connected these separate things is done by means of connectors which reflect the transferring action through one route to another. A multi - modal network model is built using this principle upon its TFL platform.

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. https://doi.org/10.1109/ICWR.2017.7959324

[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. https://doi.org/10.1109/PST.2016.7906975

[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] B.Pandu Ranga Raju, B.Vijaya Lakshmi, C.V.Lakshmi Narayana, **Detection of Multi-Class Website URLs Using Machine Learning Algorithms,** International Journal of Advanced Trends in Computer Science and Engineering, pp. 1704-1712 ,Volume 9, No.2, 2020. https://doi.org/10.30534/ijatcse/2020/122922020

[12] Roobaea Alroobaea, **An Empirical combination of Machine Learning models to Enhance author profiling performance**, International Journal of Advanced Trends in Computer Science and Engineering, pp. 2130-2137,Volume 9, No.2, 2020. https://doi.org/10.30534/ijatcse/2020/187922020