International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor (2018): 7.426

Intrusion Detection System using Ensemble Learning W-AODE and REPTree Algorithm Accuracy Graphs on WEKA

Apoorvi Nagar[1], Kapil Sharma[2]

[1, 2]Department of Computer Science, ITM University, Gwalior (M.P), India

Abstract: ***With the advancement in information and communication technology (ICT), it has become a vital component of human's life. But this technology has brought a lot of threats in cyber world. These threats increase the chances of network vulnerabilities to attack the system in the network. To avoid these attacks there are various methods in which one is Intrusion Detection System (IDS). In IDS, there are various methods used in data mining and existing technique is not strong enough to detect the attack proficiently. Weighted Average One-Dependence Estimator (WAODE) is an enhanced version of AODE and in this technique; we have to assign weights to each attribute. The dependent attributes having lesser weights by defining the degree of the dependencies. This paper deals with a novel ensemble classifier (WAODE+ RepTree) for intrusion detection system. Proposed ensemble classifier is built using two well-known algorithms WAODE and RepTree. This tree improves accuracy and reduces the error rate. The performance of proposed ensemble classifier (WAODE+ RepTree) is analyzed on Kyoto data set. Proposed ensemble classifier outperforms WAODE and RepTree algorithms and efficiently classifies the network traffic as normal or malicious.***

Keywords: Intrusion Detection System, Classification, pre-processing, Weighted Average One-Dependence Estimator, RepTree, Malicious and attacker

1. Introduction

Each PC is dependably in danger for unauthorized and intrusion, be that as it may, with sensitive and private data are at a higher risk. Intrusion Detection is a key method in Information Security assumes an imperative part detecting different kinds of attacks and secures the system framework. Interruption Detection is the way network watching and investigating the occasions emerging in a PC or system framework to recognize all security issues. IDS provide three imperative security capacities; monitor, detect and react to unauthorized activities [1]. IDS monitor the tasks of firewalls, routers, management servers and documents basic to other security mechanisms. IDS can make the security management of framework by non- expert staff conceivable by giving easy to use interface. IDSs as a rule give the accompanying administrations:

- Observing and analyzing computer and additionally network system activity.
- Audit the framework designs and vulnerabilities.
- Evaluating the integrity of basic framework and data files.
- Estimating anomalous activities.

2. Function of IDS

The IDS comprise of four key capacities to be specific, data gathering, feature selection, analysis and activity, which is given in Figure 1.

a) Data collection
This module passes the data as input to the IDS. The information is recorded into a document and afterward it is analyzed. Network based IDS gathers and alters the data packets and in have based IDS gathers points of interest like utilization of the disk and procedures of the framework.

b) Feature Selection
To choose the specific component expansive information is accessible in the system and they are typically assessed for interruption. For instance, the Internet Protocol (IP) address of the source and target framework, protocol compose, header length and size could be taken as a key for intrusion [2].

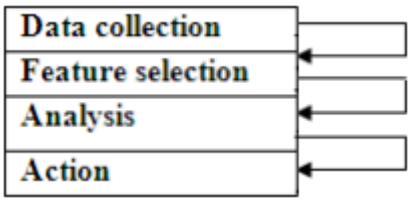

Figure 1: Functionality of IDS

c) Analysis
The data is analyzed to find the accuracy. Rule based IDS analyze the data where the moving toward traffic is checked against predefined signature or illustration [2]. Another strategy is anomaly based IDS where the framework conduct is contemplated and mathematical models are utilized to it [2].

d) Action
It characterizes about the attack and response of the framework. It can either advise the framework overseer with all the required information through email/ alarm icons or it can have a dynamic impact in the framework by dropping packets so it doesn't enter the framework or close the ports [3].

This section will discuss AODE, RF, and our proposed methodology in brief.

- **Random Forest (RF)**

In recent times Random forest, (now onwards we denote as RF) have been used widely for IDS problem. RF combines bagging and random selection of features. RF consists of many classification trees. RF improves the accuracy and reduces error rate for large data sets. RF generates out of bag error during training phase. Three Turing parameters are used in RF: No of trees, minimum node size, numbers of descriptors are used for splitting each node. Steps in Random forest are as follows

Steps in Random Forest

Step 1: From the training set, select a bootstrap sample

Step 2: On this bootstrap sample, grow an un pruned tree

Step 3: Randomly select predictors at each node and determine the best split.

Step 4: Save tree as it is and don't apply cost complexity.

- **Average One Dependency Estimator (AODE)**

Average one dependency estimator resolved the attribute independence issue in naive bayes. As naïve bayes classifier does not consider attribute interdependency, this may affect the accuracy of the IDS. High computational overheads of naïve bayes and augmented naïve bayes are overcome by AODE. AODE is capable of accurately predicting whether network traffic is normal or anomalous. AODE achieves high accuracy by averaging the aggregation of many special tree augmented naïve bayes. AODE has a low variance and supports incremental learning and can effectively handle missing values. It has classification time of O (cn2) and has a computational complexity ofO (tn2).

- **Ensemble Learning**

Ensemble learning is a new trend in AI and data mining, in which several weak learning algorithms are combined . Idea behind ensemble classification is to exploit the strength of weak learning algorithms to obtain a robust/efficient classifier. A single IDS developed with weak learning algorithm can cover and identify limited input data and no. of attacks .Ensemble classifiers are constructed by a set of weak classifiers and decision function which combines the classification results. Majority voting is simple and efficient decision function used in many ensemble techniques.

- **RFAODE**

In this subsection, we will discuss our proposed RFAODE: A novel ensemble IDS using RF and AODE. The proposed approach operates in three stages. 1) Data preprocessing (For Kyoto Data set) 2) Training and evaluating ensemble classifier 3) Testing Prior to applying ensemble classifier, it is essential to convert the features to a format which are perceivable by the ensemble classifier. In our proposed approach, features are converted from numeric to binary. Our proposed ensemble IDS will operate on this data set for classification of network traffic data. Inter quartile range (IQR) is used to remove noise and outliers in the data set. Variability is summarized by IQR.

- **Weighted Average One-Dependence Estimator (WAODE)+REP Tree**

Weighted Average One-Dependence Estimator (WAODE) is an enhanced version of AODE and in this technique; we have to assign weights to each attribute. The dependent attributes having lesser weights by defining the degree of the dependencies. This paper deals with a novel ensemble classifier

(WAODE+ RepTree) for intrusion detection system. Proposed ensemble classifier is built using two well-known algorithms WAODE and RepTree. This tree improves accuracy and reduces the error rate. The performance of proposed ensemble classifier (WAODE+ RepTree) is analyzed on Kyoto data set

3. K-Means Clustering

K-Means algorithm is a hard partitioned clustering algorithm generally utilized because of its straightforwardness and speed. It uses Euclidean distance as the similarity measure. Hard clustering means that an item in a data set can belong to one and only one cluster at a time.

It is a clustering analysis algorithm that groups items based on their feature values into K disjoints cluster s such that the items in the same cluster have similar attributes and those in different clusters have different attributes. The calculation is connected to training datasets which may contain normal and abnormal traffic without being named already. The primary thought of this approach depends on the assumption that nor mal and abnormal traffic for m diverse clusters.

The data may likewise contain exceptions, which are the data items that are altogether different from alternate items in the cluster and subsequently don't have a place with any cluster. An outlier is found by comparing at the ranges of the data items; that is, if the radius of an data item is more prominent than a give edge, it is considered as an anomaly. In any case, this does not aggravate the K- Means Clustering process as long as the quantity of anomaly s is little. [4] K- Means Clustering Algorithm is as per the following:

a) Define the quantity of clusters K. For instance, if K=2, we accept that typical and strange activity in the preparation information for m two distinct clusters..
b) Initialize the K cluster centroids. This should be possible by arbitrarily choosing K information things from the data set.
c) Compute the separation from everything to the centroids of the whole cluster by utilizing the Euclidean separation metric which is utilized to discover the comparability between the items in data set [4].

4. Support Vector Machine

Binary classification issues can be solved utilizing SVM. A SVM maps linear algorithms into non- linear space. It utilizes an element called, kernel function, for this mapping. Kernel capacities like polynomial, radial basis function are utilized to separate the feature space by building a hyper-plane. The kernel functions can be utilized at the season of preparing of the classifiers which chooses support vectors along the surface of this capacity.

SVM classify data by utilizing these support vectors that diagram the hyper-plane in the component space. [4]. This process will involve a quadratic programming problem, and this will get a global optimal solution. Suppose we have N training data points {(x1, y1), (x2, y2), (x3, y3)... (xN, yN)}, where xi € Rd and yi € {+1, −1}. Consider a hyper-plane

described by (w, b), where w is a weight vector and b is an inclination. The classification of a new object x is done with

$$f(x) = \text{sign}(w.x + b) = \text{sign}(\sum^{N} \alpha_i y_i (x_i \cdot x) + b)$$

The training vectors xi occurs only in the form of a dot product. For each training point, there is a Lagrangian multiplier αi. The Lagrangian multiplier values αi reflects the importance of each data point. At the point when the maximal margin hyper-plane is discovered, just indicates that lie nearest the hyper-plane will have $\alpha i > 0$ and these focuses are called support vectors. Every other point will have $\alpha i = 0$. That implies just those focuses that lie nearest to the hyper-plane, give the portrayal of the hypothesis/classifier. These data points serve as support vectors. Their values can be used to give an independent boundary with regard to the reliability of the hypothesis/classifier [4].

5. Literature Survey

Valli Kumari V et al. [2017] this work demonstrates a hybrid semi-supervised machine learning technique that uses Active learning Support Vector Machine (ASVM) and Fuzzy C-Means (FCM) clustering in the design of an efficient IDS. This algorithm is tested on NSL KDD bench mark IDS data set and found to be promising [5].

Shashikant Dugad, et al. [2017] in this paper, our present a state-of-the-art solution for ship intrusion detection using image processing and Support Vector Machine (SVM). The main aim is to detect the ships, which cross over the border and secured industrial spaces. Utilizing the interworking instruments of these two strategies, we can recognize the interfering boat from the always showing signs of change sea environment. SVM can be utilized as machine learning out how to prepare the framework by presenting it to various seashore situations. Subsequently, it can be utilized as a constant security framework at seashore regions [6].

Kinan Ghanem et al. [2017] in this work, SVM is regarded as a ML system that could supplement the execution of our IDS, giving a second line of recognition to decrease the quantity of false alarms, or as an alternative detection technique. We evaluate the execution of our IDS against one-class and two-class SVMs, utilizing linear and non-linear structures. The outcomes that we exhibit demonstrate that direct two-class SVM produces exceedingly exact outcomes, and the precision of the straight one-class SVM is exceptionally practically identical, and it needn't bother with preparing datasets related with malicious data. So also, the outcomes confirm that our IDS could profit by the utilization of ML methods to build its precision while analyzing datasets involving non-homogeneous highlights [7].

Adriana-Cristina Enache et al. [2017] in this paper, our propose to conduct a comparative study of feature selection methods for intrusion detection. our focus on wrapper variants of FSM which are based on swarm intelligence algorithms. To conduct our study, we construct our FSM models based on four SI algorithms (PSO, BA, BAL and BAE) in combination with traditional classifiers (SVM, C4.5 and Naive Bayes) and use the NSL-KDD dataset for our tests and comparative analysis [8].

Guohang Yin et al. [2016] in this paper, instruction detection technology are a new generation of security technology that monitor networks or systems to avoid malicious activity and policy violation. Compared with traditional security protection measures such as firewall, instruction detection can prevent attacks both from external and internal. The SVM is a statistical learning model (SLT), which shows an extraordinary advantage when dealing with small sample. Its advantages are: (1) SVM's goal is to get the optimal solution under limited samples but not infinity samples which is the prerequisite of traditional machine learning like neural network or regression. (2) SVM has a regularization parameter to maintain a strategic distance from over-fitting and uses the piece trap to travel to high-dimensional element space to build VC dimension. This manuscript is based on the SVM to extract intrusion detection information, at the same time in order to eliminate noise caused by false alarm probability, our also combined with the context validation as a preliminary analysis, so as to achieve a novel computer network intrusion detection (NCNID) algorithm [9].

Qiuwei Yang et al. [2016] in this paper, our propose an improved particle swarm optimization algorithm ICPSO, which use chaos operator ergodicity, randomness, sensitivity to initial conditions and different qualities and the ICPSO is utilized to influence the confusion into the dormancy to weight factor parameters and The chaos is connected to the optimization of the RBF kernel work parameter g and the penalty factor C, and to enhance the merging rate and accuracy of the particle swarm optimization. The experimental results demonstrate that: in respect to the PSO-SVM algorithm and GA-SVM algorithm, ICPSO-SVM enhances the effectiveness of intrusion detection, and is a suitable intrusion detection model [10].

Ajinkya Wankhade et al. [2016] this work proposes using a DIDS model for data collection across the network and a hybrid method that classifies the network activities collected in the DIDS model as normal and abnormal. This hybrid method is a combination of popular machine learning algorithms Support Vector Machine (SVM) and Ant Colony Optimization (ACO) which is to be used on a model for DIDS. Also it can detect unseen attacks of intrusion with high detection rate with minimal misclassification. Experiments show that usage of hybrid method on the DIDS model is superior to that of SVM alone or ACO alone both in terms of run-time efficiency and detection rate [11].

6. Proposed Work

Naïve Bayes is the arithmetical algorithm for the learning method for the detection of intrusion. This algorithm used attributes and they all are independent but this independency can influence the accuracy. AODE (Averaged One Dependence Estimator) algorithm used to overcome the problem and it is further accurate than Naïve Bayes. The capacity of the AODE algorithm is to observe the network traffic by probabilities. Weighted AODE (WAODE) is an enhanced version of AODE and in this technique; we have to assign weights to each attribute. The dependent attributes having lesser weights by defining the degree of the dependencies. The pre processing has performed over the data which is required in WAODE. But this technique used

for selection of feature and it consists of 2 parts for learning such as WAODE-learning and WAODE-test. In training phase, attributes are taken into consideration for assigning weights to robust the data for training. For the classification in the process, WAODE used for every given instances for testing.

Proposed Algorithm

Step:1 Input csv file and transform it into text file
Step:2 Convert data into binary form
Step:3 Select Normal Labels as 1 and Attack Labels as 1 and 2
Step:4 Performed preprocessing over the file
Step:5 Now eliminate additional attributes from data
Step:6 Separate the data into test and training file for further processing
Step:7 Select Weighted AODE and Random Forest in vote
Step:8 Select test file for further calculation
Step:9 Average of chances evaluate
Step:10 Get the output in the form of normal or malicious in network
Step:11 Stop

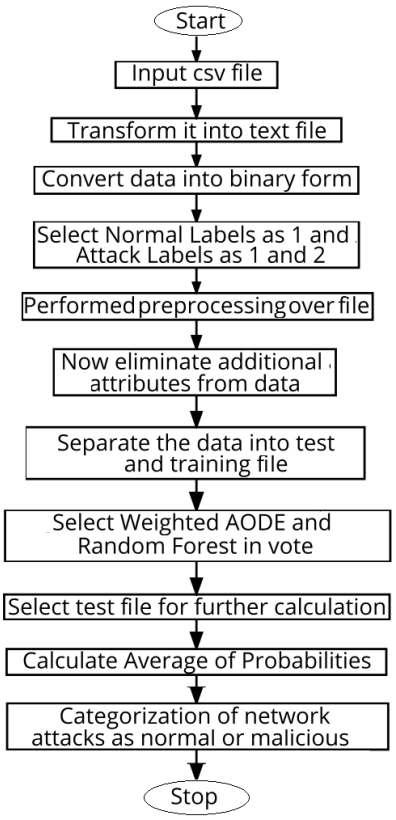

Figure 2: Flowchart of Proposed Work

7. Result Analysis

In the result analysis, the experiment of proposed work performed by using ensemble classifier. Kyoto dataset 2006 used for the investigational study of the traffic data. This dataset contains 24 features and we used only 15 features and excluded the features which are related to security analysis.

Instances: 85346
Attributes: 15
Duration_binarized
Service
SourceByte_binarized
DestinationByte_binarized
Count_binarized
Same srv rate_binarized
Serror rate_binarized
Srv serror rate_binarized
Dst host count_binarized
Dst host srv count_binarized
Dst host same src port rate_binarized
Dst host serror rate_binarized
Dst host srv serror rate_binarized
Flag
Label

Test mode: 10-fold cross-validation

1. Training

Base Work:

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.996	0.865	0.955	0.996	0.975	0.285	0.935	0.996	Attack
0.135	0.004	0.667	0.135	0.224	0.285	0.935	0.439	Normal
Weighted Avg.	0.952	0.821	0.940	0.952	0.937	0.285	0.935	0.968

=== Confusion Matrix ===

a	b	<-- classified as
80681	294	a = Attack
3783	588	b = Normal

Propose Work

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.998	0.737	0.962	0.998	0.979	0.460	0.938	0.996	Attack
0.263	0.002	0.852	0.263	0.402	0.460	0.938	0.502	Normal
Weighted Avg.	0.960	0.700	0.956	0.960	0.950	0.460	0.938	0.971

=== Confusion Matrix ===

a	b	<-- classified as
80776	199	a = Attack
3222	1149	b = Normal

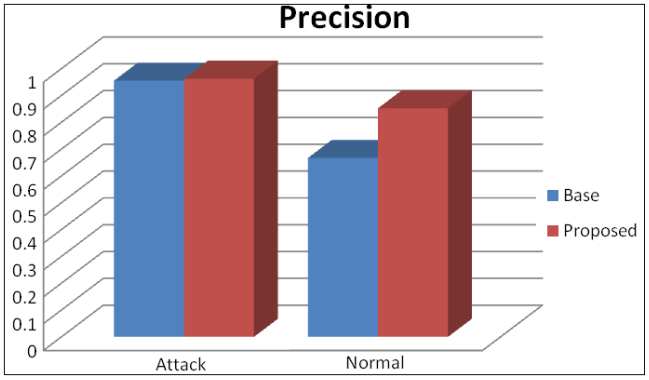

Figure 3: Precision Comparison

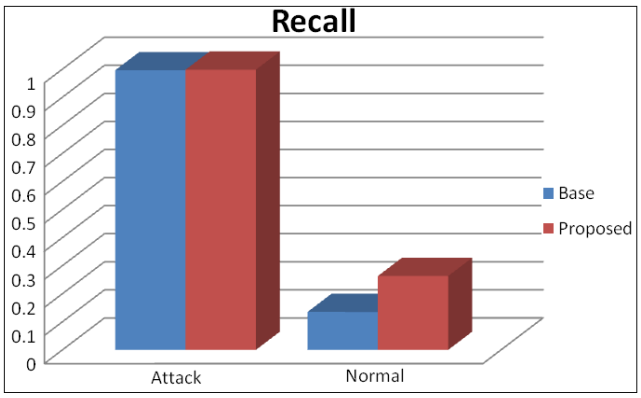

Figure 4: Recall Comparison

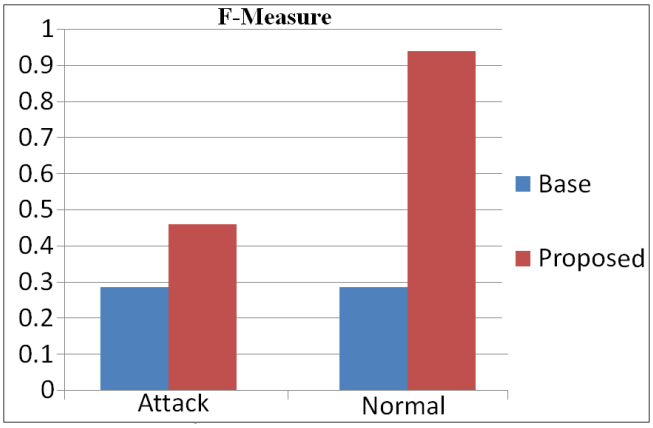

Figure 5: F-Measure Comparision

8. Conclusion

In this research paper, we proposed a novel ensemble classifier (WAODE RepTree) for intrusion detection system. The proposed approach efficiently classifies network traffic as normal or malicious. The results indicate that proposed classifier is accurate than RF and AODE classifiers. We considered Kyoto data set for experimental analysis. As Base classifiers are not capable of detecting the attacks accurately, proposed Ensemble classifier outperforms base classifiers WAODE and RepTree. The results presented in this paper show that integration of WAODE, RepTree and pre-processing technique will yield the good result for IDS.

References

[1] V. Jaiganesh, S. Mangayarkarasi , Dr. P. Sumathi "Intrusion Detection Systems: A Survey and Analysis of Classification Techniques" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013.

[2] Dr. S.Vijayarani1 and Ms. Maria Sylviaa.S "INTRUSION DETECTION SYSTEM – A STUDY" International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 4, No 1, February 2015.

[3] Sriram Sundar Rajan, Vijaya Krishna Cherukuri-"An Overview of Intrusion Detection Systems" 2013.

[4] Miss.Kavita Patond, Prof. Pranjali Deshmukh "Survey on Data Mining Techniques for Intrusion Detection System" International Journal of Research Studies in Science, Engineering and Technology [IJRSSET] Volume 1, Issue 1, April 2014, PP 93-97.

[5] Valli Kumari V, Ravi Kiran Varma P "A semi-supervised Intrusion Detection System using active learning SVM and fuzzy c-means clustering" International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017).

[6] Shashikant Dugad, Vijayalakshmi Puliyadi, Heet Palod, Nidhi Johnson, Simran Rajput, Swapna Johnny "Ship Intrusion Detection Security System Using Image Processing & SVM" 2017 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017).

[7] Kinan Ghanem, Francisco J. Aparicio-Navarro, Konstantinos G. Kyriakopoulos, Sangarapillai Lambotharan, Jonathon A. Chambers "Support Vector Machine for Network Intrusion and Cyber-Attack Detection"2017, IEEE.

[8] Adriana-Cristina Enache, Valentin Sgârciu and Mihai Togan "Comparative Study on Feature Selection Methods rooted in Swarm Intelligence for Intrusion Detection" 2017 21st International Conference on Control Systems and Computer Science.

[9] Guohang Yin, Youran Zhang and Ziyi Zhao "A Novel Computer Network intrusion Detection Algorithm Based on OSVM and Context Validation" 978-1-5090-3484-0/16/$31.00 ©2016 IEEE.

[10] Qiuwei Yang, Hongjuan Fu and Ting Zhu "An Optimization Method for Parameters of SVM in Network Intrusion Detection System" 2016 International

Conference on Distributed Computing in Sensor Systems, IEEE.

[11] Ajinkya Wankhade, K. Chandrasekaran "Distributed-Intrusion Detection System using combination of Ant Colony Optimization (ACO) and Support Vector Machine (SVM)" 2016 International Conference on Micro-Electronics and Telecommunication Engineering.