Comparative Analysis for Detecting Malicious packets in LAN Network using Bagging and Boosting Techniques

[1]Mrs. A.Vinitha, [2]Dr. B.Rosiline Jeetha

[1]M.Sc., M.Phil., (Ph.D), Assistant Professor cum Doctoral Research Scholar, Dr.N.G.P. Arts and Science College Coimbatore

[2]M.C.A., M.Phil., Ph.D. Professor & Head, Doctoral Research Supervisor, Dr. N.G.P. Arts and Science College Coimbatore.

Abstract

The security is the important one in sharing the data through the network. The intrusion detection system helps to classify the data as the normal or anomaly by using the machine learning algorithms. There is the large pool of machine learning algorithms which helps to find the anomaly. The algorithms have to find the connection that is normal or abnormal. The attacks are classified into four categories. Using the machine learning algorithm, the prediction is done to classify the packet as the normal or anomaly. The dataset used in this paper is communication of the LAN network which is happened in the very sensitive environments like military and Airforce. Various models are used to predict the anomaly but the bagging performs well in predicting the anomaly. There are various features associated in the dataset which helps to classify the network connection. The algorithms like adaboost was done with 10 iteration and the time for execution is 1.55 seconds and the accuracy are 94 %, bagging the time taken to execute the algorithm is 2.27 seconds, accuracy is 99%. Among these two algorithms the bagging performs well in predicting the abnormal data. Totally 42 attributes have been taken for the training and the testing purposes.

Keywords: *Anomaly, Adaboost, Bagging*

1. Introduction

The digital communication is very essential in the today's world. There is no time for the people to reach and receive the information directly. So, the network which helps to satisfy the needs of the user. The network connection is very important for transferring the data. There are many people which performs the hacking process for developing their company. The intrusion monitors the network and helps to identify the malicious packet. The intrusion software helps to detect the unauthorized who is trying to access our own resources. There are various attacks are there. The usage of IOT services also increases the devices in the network and it also increases the attacks in many ways. Nowadays the developed work depends on how safely that are transferred through the network. So, to provide proper security there should be system like firewall to watch like a real time in the network. If there is any change the system should give some clue to find the intrusion.

The machine learning algorithms plays a major role in various fields. The machine is initially trained to classify the data. The machine learning algorithm will update automatically and brings the better results. The machine learning algorithm can be classified into supervised, unsupervised, and reinforcement learning.

There are 42 features in the data set which is used to train the algorithms and to perform the classification. In this Ada Boost algorithm, bagging is used because the data set is very huge. In this algorithm the weak entities are predicted to form the strong entity. This is mostly applied for the massive data. This algorithm will support in implementing in R and the python language easily. The bagging is one of the ensemble-based approach which helps to provide high accuracy and all the individual models are combined together to create the better results. This algorithm also avoids the overfitting of data. It helps to average all the models. The ensemble-based model helps to combine the weak entities together.

2. Literature Survey

The author Anish Halimaa et al [1] states that the security is the very important that each and every individual tries to achieve in this digital world. There are many competitions and the compromises in the today's business life cycle. The intrusion detection occurs in all places. The intrusion detection is the identification of the malicious packet which alerts the network administrator or the user. The intrusion detection can be done at two ways. They are host level and also at the network level. There are two types of intrusion detection. They are anomaly detection and the misuse detection. The anomaly detection can be done observing the abnormal patterns that is found in the network. The misuse detection is done by storing the attack signatures in the databases and there by comparing the results. The author has used the classification methodology which will help to monitor the network. In this paper SVM and the naïve bayes are used to find the malicious code in the network. The author mainly concentrates on the finding of accuracy and increasing it. Another one task is to find the misclassification.

The intrusion creates the harm for the system and it also spoils the sensitive information. The author concentrates on the reducing the false alarm rate. Initially the data set will undergo the pre-processing for removing the non-numeric values and the symbolic values. There are different types of attacks are there. Day by day the attacker is very active and finding new ways and the technics. Two algorithms from the machine learning are compared together. In these 19000 instances are collected and it is compared. The SVM and the Naïve are compared together in various ways. The SVM shows the accuracy about the 97.29 and the Naïve shows the 67.26. Initially the machine learning algorithms are trained and later on the testing process will be done. The misclassification rate between the two algorithms are also calculated. In future the author suggests the hybrid classification and the other types of attacks can be predicted.

The intrusion detection can be done at different levels. Based on the organization needs the intrusion detection system can be selected to monitor the network.

The author GP Gaikwad [2] et al states that the intrusion detection is the very important for protecting the information from many threats and the vulnerabilities. There are many machine learning algorithms are used to predict the results. In this paper the ensemble-based approach with the decision tree algorithm as the base model is used to detect the intrusion by reducing the false alarm rate and increasing the accuracy. The bagging is the best suited algorithm for providing the good results. The algorithm provides the good results by giving the less false positives. The network intrusion detection occurs in all areas. The model takes the very less technique and used the cross-validation technique. The model is compared with other models and it provides the good accuracy with good results.

The Mohammad Mahmood [3] Otoom et al states that the day by day there is an increase in the artificial intelligence and many industries generates the huge amount of data in the network. So, there is the very big challenge for the cyber crime to protect the systems. There is no possibility of developing the network without any cybercrimes. So, there must be good machine learning algorithm to predict his vulnerability and to inform the user to secure the data. The author used the bagging algorithm to reduce the false alarm rate and to increase the accuracy. The author has used the J48 as the base classifier in the bagging method. After evaluating and testing the results the author says the J48 outperforms well when it is compared with other algorithms. The network security is the very much challenging task that happens regularly in all time. The hacking of the personal information unauthorised will make the others to loss of information.

The author Bayu Adhi Tama [4] et al states that the intrusion detection can be done at various levels. The author has collected various datasets and the various areas where the results are analysed in various categories. The models are developed to predict the data in various ways. So, the very small companies will not think about spending a large amount. Some algorithms are examined in the statistical way where the accuracy, ROC curve is examined. In this paper different types of data set are examined and various results have been generated.

Ying Zhou [5] et al states that the intrusion detection is done with the 802.11 wireless network. The Ada boosting method is used to achieve good results. The author also used the optimization algorithm to achieve good results. The different types of data sets are compared to produce the good results.

The author [6] et al Maya Hilda Lestari Louk et al states that there are many existing algorithms for detecting the intrusion but even though they are suffering from the false alarm rate. So, the bagging and the gradient boosting algorithm are used to predict the anomaly and to test with the different available data sets. The author states that this is the best algorithm when it is compared with other algorithms also.

Iwan Syarif [7] et al states that the combination of the ensemble-based approach will help to detect the intrusion in the network. The combination of bagging, boosting and the stacking are combined together to predict the intrusion in many ways. This combination will help to improve the accuracy and to reduce the false positive rate. Three different types of the algorithms are choosed as the base classifier.

Owais Bukhari [8] et al states that the introduction of IOT enables the various different kinds of attacks and it leads to loss of th sensitive information. The anomaly detection is very important because many abnormal patterns are found to predict it. The IOT is the important nervous system for all the IOT application where the smart cities are implemented. The base classifiers are selected and combined with the ensemble-based methods which will show the good results when it is compared with other methods.

The author [9] et al states that intrusion can be predicted by using the Ada boost algorithm. The traditional algorithm shows that it is not enough to increase its ability to find the accuracy. So, the author made some changes in the weak learners and also in weightage of the weak learners. This will help to increase the accuracy in good way. So, variations in the threshold value will give the better results. The author [10] et al states tells that the security is the big issues that has to be achieved for the very big organization. Many users are easily compromising the network to access the data in an unauthorised manner. So, the convolutional neural network is combined with the bagging method to create deep bagging convolutional neural network. The process in this algorithm is pre-processing, feature selection and the classification is done to provide good results.

3. Problem Statement

The intrusion detection is done to predict the malicious in the network and to alert the user by providing some clues. There are various machine learning models which helps to perform this work. The data set used in this is LAN network data which is highly sensitive when it is used in the military and Airforce. The security is vey much essential in these areas. The meta heuristics algorithms are used to predict the malicious and it will help to combine the various models which helps to aggregate and give best results. These algorithms will have the capability to deal with very large data sets.

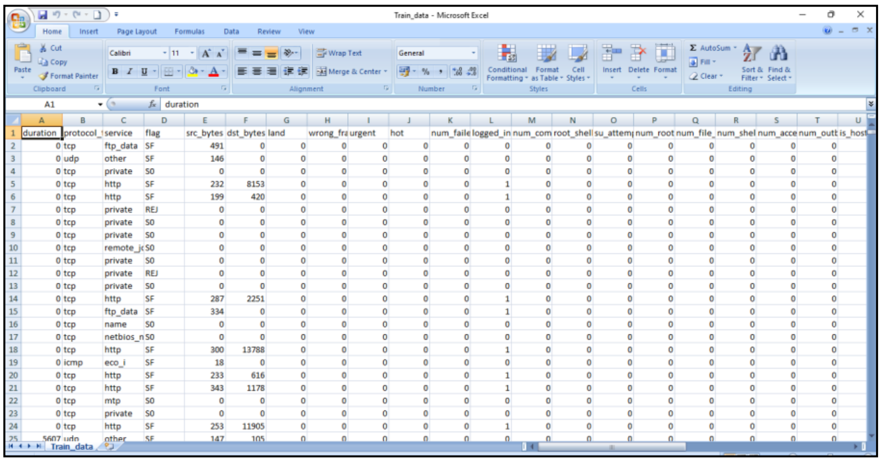
3.1Bagging

The bagging is called the ensemble-based method which combines all the weak learners in the model. It is also called as the bootstrap aggregation. It has the capability of applying in the high dimensionality of data. It also helps to reduce the variance. The algorithm will undergo several iterations to combine all weak learners. The sampling process is done in the data set and the weak learners are combined to create the strong learners. It can be performed with number of iterations. It has the capability of performing the execution in parallel manner.

The bagging can be applied in many different areas. Each created sample will undergo some weightage and finally it will be combined to give good results.

Model Training

The selected subsets of the data set with replacement are trained and all the individual predictions are combined together to make single prediction. There are various advantages while working with the bagging technique. There are 42 attributes that are used for the training purposes. The data set as follows



duration	protocol_	service	flag	src_bytes	dst_bytes	land	wrong_fra	urgent	hot	num_faile	logged_in	num_com	root_shell	su_attem	num_root	num_file_	num_shel	num_acce	num_outb	is_host
0	tcp	ftp_data	SF	491	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	udp	other	SF	146	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	http	SF	232	8153	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
0	tcp	http	SF	199	420	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	remote_jc	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	http	SF	287	2251	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
0	tcp	ftp_data	SF	334	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
0	tcp	name	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	netbios_n	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	http	SF	300	13788	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
0	icmp	eco_i	SF	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	http	SF	233	616	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
0	tcp	http	SF	343	1178	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
0	tcp	mtp	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	tcp	http	SF	253	11905	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
5607	udp	other	SF	147	105	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Table 1

10-Fold Cross Validation

The model will undergo some cross-validation technique. There are some chances of occurring the overfitting of the data. to overcome the problem of the overfitting the cross-validation process is done. The entire data set is divided into ten set and the k value is 10. For each iteration the one part will be used for the testing and the remaining will be used for the training purposes. There are different validation techniques which will help to overcome from the problem of overfitting. The overfitting is nothing but the model will give good results for the training set and there will be poor performance for the validation and the testing purposes. To overcome this problem 10-fold process is done to overcome from the bias.

The reduced error pruning algorithm is used as one of the decision tree algorithms which will be used to create the decision tree. The time taken to build the model is 2.27 seconds.

Stratified cross validation

The stratified cross validation technique is used to divide the data set equally from each class so to increase the accuracy and to avoid the overfitting of the data. The results are as follows.

Correctly Classified Instances	**25093**	**99.607%**
Incorrectly Classified Instances	**99**	**0.393%**
Kappa statistic	**0.9921**	
Mean Absolute error	**0.0066**	
Root mean Squared error	**0.0555**	
Relative absolute error	**1.3318%**	

Root relative squared error	**11.121%**
Total number of Instances	**25192**

Table 2

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.997	0.005	0.995	0.997	0.996	0.992	0.999	0.999	Normal
0.995	0.003	0.997	0.995	0.996	0.992	0.999	0.999	Anomaly
0.996	0.004	0.996	0.996	0.996	0.992	0.999	0.999	Weighted Average

Table 3

The confusion matrix is used to analyse the performance of the classification algorithms. The class A is normal and the class B is anomaly.

A	B	Class
13413	36	A- Normal
63	11680	B- Anomaly

Table 4

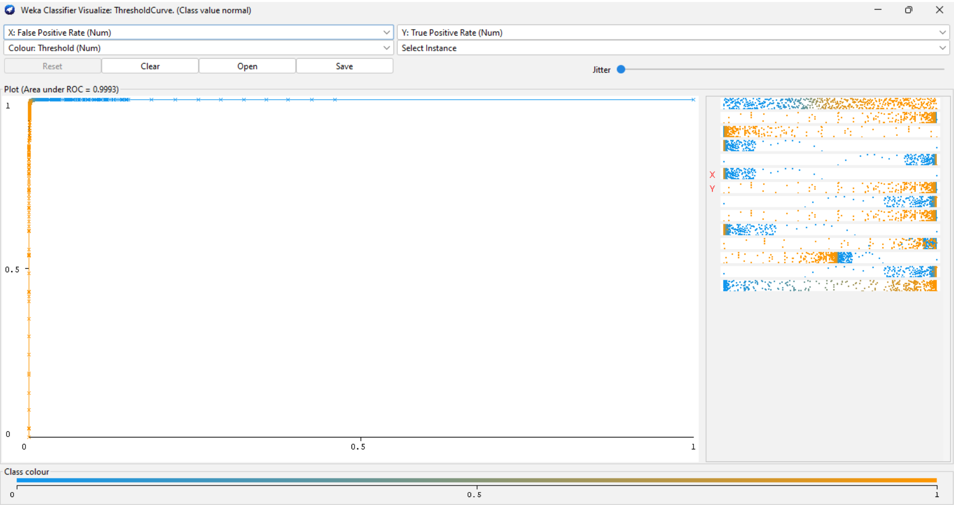

Weka Classifier Visualize: ThresholdCurve. (Class value normal)
X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot (Area under ROC = 0.9993)
1
0.5
0
0
0.5
1
X
Y
Class colour
0
0.5
1

Figure 1

The above diagram shows the graph of the confusion matrix in the form of visual representation.

3.2 AdaBoost

The boosting is the one of the ensemble-based method which will be used to predict the data. Instead of dividing the data set into multiple sets the data set is given to one model with assigned weights. In that some of them will provide the correct results and some of the data will not provide the correct results. The incorrect results are given to the next model with increase in weightage. The second model may produce some incorrect results for some data. These incorrect results are again given to the third model and this process repeats until all are correctly classified or we have to fix some threshold value which after reaching that we can stop the process.

In this 10-fold cross validation is done. The algorithm will go for many iterations. Then the results are as follows.

Correctly Classified Instances	**23773**	**94.3673%**
Incorrectly Classified Instances	**1419**	**5.6327%**
Kappa statistic	**0.8866**	
Mean Absolute error	**0.0796**	
Root mean Squared error	**0.1949**	
Relative absolute error	**16.0005%**	
Root relative squared error	**39.0706%**	
Total number of Instances	**25192**	

Table 5

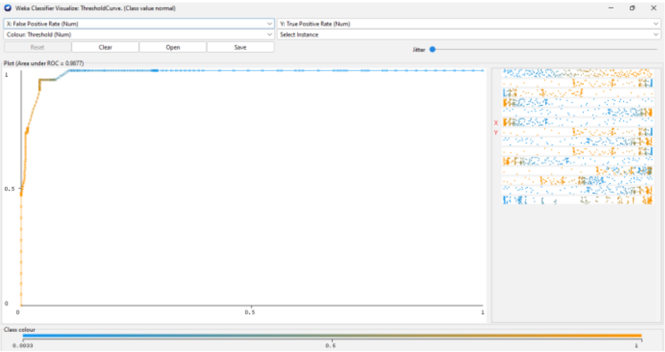
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.960	0.075	0.936	0.960	0.948	0.887	0.988	0.987	Normal
0.925	0.040	0.953	0.925	0.939	0.887	0.988	0.989	Anomaly
0.944	0.059	0.944	0.944	0.944	0.887	0.988	0.988	Weighted Average

Table 6

A	B	Class
12908	541	A- Normal
878	10865	B- Anomaly

Table 7

Totally ten iterations are performed. The time to build the model is 1.55 seconds.



Weka Classifier Visualize: ThresholdCurve. (Class value normal)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.9877)

Figure 2

4. Comparative Analysis between the Algorithms

In this the bagging and the boosting algorithm are compared. In this by comparing these two methods the bagging performs well and its accuracy is high when it is compared to boosting.

5. Conclusion

The security is the very important to share the data in the network. Many of them still searching to implement the good security. Many hackers are trying to access the confidential information in all ways. So, there should be an alert to secure our information. The IDS will support in providing the security. The introduction of artificial intelligence and the connection of many devices in the network makes the severe in achieving the security. In this paper the ensemble-based approaches are compared and the bagging performs well for the LAN network. In future different types of data sets are tested in these methods.

References

[1] Anish Halimaa A, Dr. K. Sundarakantham, “Machine Learning Based Intrusion Detection System “, IEEE Explore 2019.

[2] D.P. Gaikwad , Ravindra C. Thool, “ Intrusion Detection System Using Bagging Ensemble Method of Machine Learning” , ICCCCA, 2015.

[3] Mohammad Mahmood Otoom , Khalid Nazim Abdul Sattar, Mutasim AI Sadig, “ Ensemble Model for Network Intrusion Detection System Based on Bagging Using J48” , Advances in Science and Technology Research Journal.

[4] Bayu Adhi Tama, Sunghoon Lim, “Ensemble Learning for Intrusion Detection Systems: A Systematic Mapping Study and Cross Bench mark Evaluation”, Elsevier 2021.

[5] Ying Zhou, Thomas A, Mazzuchi, Shahram Sarkani, “M -Ada Boost – A Based Ensemble System for network intrusion detection”, Elsevier 2020.

[6] Maya Hilda Lestari Louk, Bayu Adhi Tama, “ Dual – IDS: A Bagging Based Gradient Boosting Decision Tree Model for Network Anomaly Intrusion Detection System”, Elsevier 2021.

[7] Yi Ding, Hongyang Zhu, Ruyun Chen and Ronghui Li, “ An Efficient Ada Boost Algorithm with the Multiple Thresholds Classification”, Applied Sciences MDPI, 2022.

[8] Owais Bukhari, Parul Agarwal, Deepika Koundal, Sherin Zafar, “ Anomaly detection using Ensemble Techniques for Boosting the Security of the Intrusion Detection System” , Elsevier, 2023.

[9] Mathiyalagan Ramasamy, Pamela Vinitha Eric, “An Improved Deep Bagging Convolutional Neural Network Classifier for Efficient Intrusion Detection System”, Bulletin of Electrical Engineering and Informatics, 2022.

[10] Hui Zhao, “ Intrusion Detection Ensemble Algorithm based on Bagging and Neighbourhood Rough Set”, International Journal of Security and its Applications, 2018.