EVALUATION OF VARIOUS CLASSIFIERS FOR NSL-KDD DATASET USING MACHINE LEARNING APPROACH

[1]Tanya Garg, [2]Sukhdeep Kaur

[1]Assistant Professor,[2]Assistant Professor
[1]Department of Computer Science& Engineering
[1]Gulzar Group of Institutes, Ludhiana, India

Abstract : As the traffic on the network is increasing day by day due to intense use of internet, security is considered as the main issue. The network security procedures need an extreme attention to analyze the network traffic in an efficient manner. The various new evolving intrusions affect the network security adversely. There are number of security tools developed to prevent the network from the various types of intrusions but the rapid rise of intrusion activities is a concerned issue. An Intrusion detection system is used to analyze the network traffic for intrusions and classifies the network traffic as normal or an attack. Classification methods assist to design "Intrusion Detection Models" which can distinguish the normal network traffic and intrusive traffic. In this paper we are going to analyze the various classifiers that can be used to design the Intelligent Intrusion Detection Model using machine learning methodology. These classifiers have been evaluated using the popular and most effective Data Mining Tool that is called as WEKA (Waikato Environment for Knowledge Analysis) using all the attributes of the NSL-KDD Dataset. The experiments have been performed by taking the instances of Training and Testing Datasets. The statistical techniques have been used to identify the best classifier among all classifiers. Accuracy, True Positive Rate (TPR), Precision performance metrics have been considered for evaluating the best classification algorithms.

Index Terms – **Classifiers, Data Mining, NSL-KDD Dataset, WEKA, Machine Learning, statistical techniques, Intrusion Detection System, Performance Metrics.**

I. INTRODUCTION

Due to increase in amount of data on the network, our important data is becoming more vulnerable to the malicious attacks. So it is very important to analyze those network data to protect our important data from any suspicious activity. Intrusion detection is very important to detect that network traffic for which specialized hardware and software systems known as network intrusion detections systems commonly called as IDS are used. They have the capability to differentiate network traffic and filter the data. A number of Machine learning algo0rithms can be used to distinguish the network traffic into intrusion and normal. This is basically called as Intrusion Detection Model.The performance of this intrusion detection model varies according to some performance metrics. In this work, NSL-KDDCup Dataset has been used to train and test the model to design the intrusion detection model. Several Statistical techniques have been used to rank the performance of various intrusion detection models. I this paper, initially, machine learning algorithms and data mining tool that has been used is introduced. After that the dataset used for experiments and the technique followed to evaluate the performance of algorithms on the basis of certain performance metrics has been discussed.

II. MACHINE LEARNING TOOL (WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS)

The tool that has been used in this work is WEKA,[1] a data mining tool that uses machine learning approach to analyze and design the Detection model and its performance. It offers various types of mechanisms but the mechanism we used is Classification. This tool is a collection of various types of algorithms and the code is written in Java language. It has about 50 preprocessing tools, 80classification algorithms and attributes evaluators for feature selection process [2]. WEKA supports the data in ARFF Format (Attribute Relation File Format).The algorithms can be applied directly on the chosen dataset to evaluate the performance [7].

III. MACHINE LEARNING ALGORITHMS

Machine learning algorithms are also called as classifiers or classification algorithms. They are used to train and test the data model. They are used to classify the network traffic when used on dataset. They are responsible for classification of network traffic as normal or intrusive, if intrusive then to which category it belongs. In our work there are 80 classifiers that are compatible on our chosen dataset and on the basis of their performance we have evaluated best machine learning algorithms and then these classifiers have been used for the selection of best feature sets. Some of the classification algorithms used are discussed in this section:

Random Forest: It is an ensembling classification algorithm and is based upon decision tree algorithm and produces output in the form of individual trees. This algorithm is a combination of bagging idea and random selection of features to construct a collection of decision trees with controlled variation. It is one of the highest accurate classifier for many datasets. A large number of variables can be handled by this algorithm without ignoring any variable [22].

Random Tree: Random Tree as its name indicating it's a tree build by picking random branches from a possible set of trees. Each tree has an equal probability of being get sampled in this algorithm or we can say the trees are distributed in a uniform way. Random Trees can be generated easily and efficiently. Combination of large sets of Random trees mostly designs accurate models [17].

JRip (Extended Repeated Incremental Pruning) is that type of rule based classifier that implements a propositional rule that performs repeated pruning to reduce errors and also called as RIPPER (Repeated Incremental Pruning to produce Error Reduction). JRip is a rule learner that exactly works like commercial rule learner RIPPER [4].

NB Tree: NB Tree (Naïve Bayes) is used and applicable to scale large databases and used to improve the performance of decision trees and Naïve Bayesian Classifiers. Attribute need not to be independent for this classification algorithm [12].

Rotation Forest: Rotation Forest is also an ensemble classifier that transforms the data, subset of instances, subsets of classes and subsets of features using Principal Component Analysis because this method of transforming data requires less storage space and has low computation time. Rotation Forest is based upon two base classifiers: decision tree and Forest. It works on two key components: diversity and accuracy [5].

IV. KDD CUP DATSET

Dataset that has been used for all the experimental process has been discussed here in this section, It consists of randomly selected instances from NSL-KDD Dataset [5]. It classifies network traffic in five categories. The number of each class instances included in training and testing dataset are mentioned in table-1

Table 1. Description of Dataset

Class Type	Instances in Training Dataset	Instances in Testing Dataset
Normal	44500	1200
Dos	10000	20000
Probe	3000	5800
U2R	12	20
R2L	30	500

V. PERFORMANCE METRICS

The performance metrics that have been used to evaluate the performance are discussed as below:

Accuracy: It is the percentage of correct predictions. On the basis of Confusion Matrix it is calculated by using the formula below:

Accuracy= TP+TN/n Here n is total number of instances.

TPR: True Positive Rate is same as accuracy so we have not considered this metrics.

Precision: It is a measure which estimates the probability that a positive prediction is correct

Precision=TP/TP+FP

VI. RESULTS AND DISCUSSION

All the experiments have been evaluated on selected instances from NSL-KDD Dataset using various machine learning algorithms using 10 cross validations. These algorithms have been evaluated using WEKA Data mining tool. Experiments have been performed to compare the performance of combinations of ranking based feature selection techniques. Then the performance of resultant sets has been evaluated using the classification algorithms. Finally the performance of all the combinations of feature sets have been compared and using Garret's Ranking Technique, ranks have been assigned to the feature sets (3000) The top five ranked classification algorithms are mentioned in the table 2.

Table 2. Performance of top ten classification algorithms using 41 attributes

Name of Classifier	Accuracy	Recall	Precision	Rank
Rotation Forest	96.4	0.964	0.983	1
Random Tree	96.14	0.961	0.979	2
Random Committee	96.11	0.961	0.982	3
Random Forest	96.12	0.961	0.979	4
IBK	96.08	0.961	0.977	5

From the table II. It can be observed that Rotation Forest Classification Algorithm perormed best in all aspects showing highest accuracy of 96.4 % but it can be observed that there is only slight difference between the performance of all the machine learning algorithms.

VII. CONCLUSION

In this work, various feature selection techniques and classification algorithms have been reviewed. All the experiments in this work have been performed using Filter model.. Rotation Forest is the best classification algorithm having highest rank.

The performance using combinational approach of feature selection is much better than the performance of features set selected by individual feature selection techniques. At present this work has been focused only on Filter Model and few performance metrics have been considered but in future this work can be extended to evaluate another feature selection models such as Wrapper or Hybrid model using all the performance metrics.

REFERENCES

[1] R. Dash, Selection of the Best Classifier from Different Datasets Using WEKA, IJERT, Vol.2 Issue 3, March 2013.\

[2] H. Nguyen and D. Choi, Application of Data Mining to Network Intrusion Detection: Classifier Selection Model, @Springer Verlag Berlin Heidelberg, 2008.

[3] M. Panda and M. Patra, A Comparative Study of Data Mining Algorithms For Network Intrusion Detection, IEEE, First International Conference on Emerging Trends in Engineering and Technology." 2008.

[4] M. Panda and M. Patra, Ensembling Rule Based Classifiers for Detecting Network Intrusions, IEEE Conference on Advances in Recent Technologies in Communication and Computing, 2009.

[5]B. Neethu,, Classification of Intrusion Detection Dataset using machine learning Approaches, IJECSE,2013.

[6]S. Garcia and F. Herrera, An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons" , Journal of Machine Learning Research 9, 2008.

[7] M. Othman and T. Yau, Comparison of Different Classification Techniques using WEKA for Breast Cancer, 2012.

[8]Kdd cup 99 intrusion detection data set..Online Available:http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[9] L. Kuncheva, "Combining Pattern Classifiers: Methods and Algorithms (Kuncheva, LI; 2004) [book review]." Neural Networks, IEEE Transactions on18.3 (2007): 964-964.

[10] P. Sangkatsanee, N. Wattanapongsakorn and C. Charnsripinyo. "Practicle Real- time intrusion detection using Machine learning approaches." Computer Communications (2011): 2227–2235.

[11] S. Yang, K. Chi Chang , H. Wei and C. Lin. "Feature weighting and selection for a real-time network intrusion detection system based on GA with KNN." Intelligence and Security Informatics (2008): 195-204.

[12]S. Mukherjee and N. Sharma "Intrusion Detection using Naive Bayes Classifier with Feature Reduction" Procedia Technology 4 (2012) 119 – 128

[13] Sajja, A. A, Knowledge Based Systems. Jones and Bartlett, 2012.

[14]Rich, E., Artificial Intelligence (3rd Ed.). Tata McGraw Hill, 2013.

[15] Pang-Ning, T. V., Introduction to Data Mining. Pearson. 2013.

[16] Alpaydin, E., Introduction to Machine Learning (2nd Ed.). PHI, 2010.

[17]Flach, P., Machine Learning the Art and Science of Algorithms that Make sense of Data. Cambridge, 2012.

[18] Jiawei Han, M. K., Data Mining Techniques and Concepts (3rd Ed.). Morgan Kauffman, 2013.

[19] Ian H. Witten, F. E., Practical Machine Learning Tools and Techniques (2nd ed.), 2012.

[20]V. Labatut and H. Cherifi, Evaluation of Performance Measurse for Classifiers Performance

[21] Liu H, Motoda H, Setiono R. & Zhao Z., Feature Selection: An Eve Evolving Frontier in Data Mining", JMLR: Workshop and Conference Proceedings Vol.4, *Publisher:*Citeseer, pages 4-13, 2010.

[22]Vege, Sri, H., Ensemble of Feature Selection Techniques for High Dimensional Data (Published Master's Thesis). Western Kentucky University, 2010.

[23] Y. Wang, F. Makedon, Application of ReliefF feature filtering algorithm to selecting informative genes for cancer classification using microarray data, Computational systems bioinformatics conference, 2004 IEEE, pages 497 – 498.

[24] S. Pulatova, Covering (rule-based) algorithms Lecture Notes in Data Mining., World Scientific publishing Co, pages 87-97, 2006.

[25]D. Ienco, R. G. Pensa, R. Meo, Context-based Distance Learning for Categorical Data clustering, LNCS 5772, Springer, Berlin, pages 83 – 94, 2009.

[26] T. Garg and S.S. Khurana, Comparison of Classification Techniques for Intrusion Detection Dataset Using WEKA, Proceedings IEEE, International Conference on Recent Advances and Innovations in Engineering, 2014.