A Comparative Study of Application of Classification Algorithms on KDDCup Dataset to Detect Intrusions using WEKA Tool

S. Venkata lakshmi
Department of Computer Science,
Loyola College,
Chennai, India.

Dr. T. Edwin Prabakaran
Department of Statistics,
Loyola College,
Chennai, India.

***Abstract*—We have networks ubiquitous now-a-days. Networks offer varied advantages to the users like resource sharing, cheaper cost and time saving etc., In spite of the many advantages present, networks are prone to various types of attacks. Networks are safeguarded by various means like firewall, Intrusion detection system, etc., Intrusion detection system is a system which monitors the activities in and around the system and raises an alert in case an abnormal or suspected malicious activity occurs. Intrusions can be detected in systems using various algorithms like k-means, k-medoids, etc., In this paper the effect of various classification algorithms like BayesNet, NaiveBayes, Randomforest and SimpleCart on 20%KDDCup dataset is being analysed using WEKA tool.**

***Keywords*— Networks, Intrusion detection, Classification Algorithms, KDDCup dataset, WEKA tool**

I. INTRODUCTION

Intrusion is defined as any set of action that can compromise the integrity, confidentiality and availability of system resources [1]. An intrusion attempt or threat is the potential possibility of a deliberate unauthorized attempt to access information, manipulate information or render a system unreliable or unusable [2]. Two types of intrusion detection exists namely misuse detection and anomaly detection [6]. Misuse detection refers to the detection of well known patterns and anomaly detection refers to the detection of unknown patterns [7][8][11].

The paper is organized as follows: Section 2 gives the introduction to the intrusion dataset used. Section 3 gives a brief explanation of the work done in the previous paper and Section 4 focuses on the introduction of WEKA tool. Section 5 represents the application of various classification algorithms with 20%KDDCup dataset using WEKA tool and the comparative plot. Section 6 presents the conclusion.

II. INTRUSION DATASET

The **KDD Cup dataset [3]** is used for intrusion detection problems and is also used in this paper. This dataset is considered to be the benchmark data in Intrusion detection. The dataset was a collection of simulated raw TCP dump data over a period of nine weeks on a local area network. The known attack types are those present in the training dataset while the novel attacks are the additional attacks which are not present in the training dataset [1][3]. There are various attacks like Buffer overflow, Perl, Portsweep, Neptune , Smurf, Teardrop, Guess password, IPSweep etc., The simulated attacks fall in one of the following categories: Denial of Service attack, User to Root attack, Remote to Local attack and Probing attack. The training dataset consists of 4,94,021 records. The testing dataset consists of 3,11,029 records. In each connection record there are 41 attributes describing different features of the connection. They are duration, protocol_type, service, flag, src_bytes, dst_bytes, land, wrong_fragment, urgent, hot, num_failed_logins, logged_in, num_compromised, root_shell, etc [3]., In this paper, 20%kddtestdata is taken in .arff format.

III. RELATED WORK

In our previous paper [9], sample KDD Cup dataset was taken and k-Nearest Neighbour algorithm was applied with whole set of attributes and also on the dataset with 5 different feature subsets of attributes. It was identified that the feature set (FS5) consisting of only 7 attributes yielded the best results in terms of identifying the maximum number of attacks [4][5][9].

IV. INTRODUCTION TO WEKA TOOL

WEKA stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, New Zealand. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be directly applied to the dataset or called from java code. Weka contains tools for data preprocessing, classification, regression, clustering, association rules and visualization [10]. It is well suited for developing new machine learning schemes. The dataset used in Weka is to be in the .ARFF format. This type of file consists of a header which describes the attribute types and a data section which is a comma separated list of data.

V. RESULTS

In this paper, Weka tool is used to apply BayesNet, NaiveBayes, Randomforest, Simplecart classifiers to 20%KDDCup testing dataset. There were totally 2098 instances in the data set. Upon applying BayesNet classifier, 1986 were correctly classified instances and 112 were incorrectly classified instances. The confusion matrix for the above application is represented as follows:

Table 5.1
Confusion Matrix for BayesNet Classifier

Normal	Anomaly
885	48
64	1101

Upon applying NaiveBayes classifier to the same dataset, the correctly classified instances were 1729 and incorrectly classified instances were 369.

Table 5.2
Confusion Matrix for NaiveBayes Classifier

Normal	Anomaly
884	49
320	845

Upon applying RandomForest classifier to the same dataset, the number of correctly classified instances were 2094 out of 2098 total instances in the dataset. Only 4 instances were incorrectly classified instances.

Table 5.3
Confusion Matrix for RandomForest Classifier

Normal	Anomaly
932	1
3	1162

Finally, on applying simplecart classifier, 2075 instances were correctly classified whereas 23 were incorrectly classified instances.

Table 5.4
Confusion Matrix for SimpleCart Classifier

Normal	Anomaly
917	16
7	1158

Table 5.5 Comparison of all four classifiers

Classification method	Correctly classified instances	Incorrectly classified instances
BayesNet	1986	112
NaiveBayes	1729	369
Randomforest	2094	4
Simplecart	2075	23

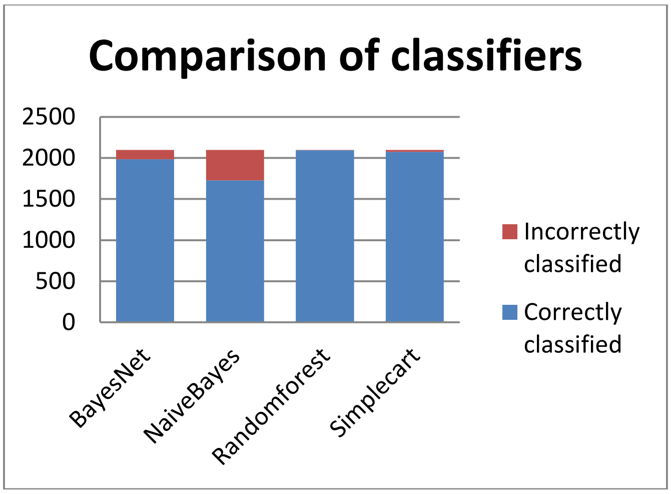

Fig 5.1 Bar diagram for comparison of classifiers

VI. CONCLUSION

In this paper the performance of the various classification methods like BayesNet, NaiveBayes, RandomForest and Simplecart on 20%KDDCup training dataset is analysed using WEKA tool. It has been identified that BayesNet method has 1986 correctly classified instances and 112 incorrectly classified instances. This shows that 94.66% instances are correctly classified. NaiveBayes method has 1729 correctly classified instances and 369 incorrectly classified instances. This shows that 82.42% instances are correctly classified and 17.58% instances are incorrectly classified. The third classification method RandomForest has 2094 correctly classified instances and only 4 incorrectly classified instances. This shows a drastic improvement in correctly classified instances which is 99.81%. Finally, on application of Simplecart classification method there are 2075 correctly classified instances and 23 incorrectly classified instances which amount to 98.91% of correctly classified instances.

From this, it is identified that the RandomForest classification method proves to be the best when compared to BayesNet, NaiveBayes and Simplecart classification methods for the given 20%KDDCup dataset. Also it is found out that Simplecart method is better than BayesNet and NaiveBayes for the given 20%KDDCup dataset. In future, the methods shall be applied with selected attributes instead of all the 41 attributes for every connection and the results could be compared and analysed. The real time data if converted into the .ARFF format shall also be used in the WEKA tool for further analysis.

VII. REFERENCES

[1] Adebayo O. Adetunmbi*, Samuel O. Falaki, Olumide S. Adewale and Boniface K. Alese, "Network Intrusion Detection based on Rough Set and k-Nearest Neighbour", International Journal of Computing and ICT Research, Vol.2,No.1,pp.60-66.,2008,http://www.ijcir.org/volume1number2/article7.pdf.

[2] J.P.Anderson, "Computer Security Threat Monitoring and Surveillance, Technical Report", Anderson Co., Fort Washington, Pennsylvania, April 1980.

[3] KDD CUP 1999 DATASET: http://kdd.ics.uci.edu/databases/kddcup99/

[4] Kok-Chin Khor, Choo-Yee Ting and Somnuk-Phon Amnuaisuk, "From Feature Selection to Building of Bayesian Classifiers: A Network Intrusion Detection Perspective", American Journal of Applied Sciences 6(11),1948-1959,2009.

[5] Neveen I.Ghali , "Feature Selection for Effective Anomaly Based Intrusion Detection", International Journal of Computer Science and Network Security, Vol.9 No.3 March 2009.

[6] Ravi Ranjan, G.Sahoo, "A new Clustering approach for anomaly intrusion detection", International Journal of Data Mining and Knowledge Management Process, Vol.4, No.2, March 2014.

[7] SANS Institute InfoSec Reading Room, "Understanding Intrusion Detection Systems".

[8] S.Ganapathy, N.Jaishankar, P.Yogesh, A.Kannan, "An Intelligent Intrusion Detection System using Outlier detection and Multiclass SVM", International Journal on Recent Trends in Engineering and Technology, Vol.5,No.1,March 2011.

[9] S.Venkata lakshmi, T.Edwin Prabakaran, "Application of k-Nearest Neighbour Classification Method for Intrusion Detection in Network Data", International Journal of Computer Applications (0975-8887) Volume 97 – No.7, July 2014.

[10] Weka Manual, http://www.ittc.ku.edu/~nivisid/WEKA_MANUAL.pdf

[11] Yang Li, "An effective TCM-KNN Scheme for High-Speed Network Anomaly Detection", International Journal of Advanced Science and Technology, Vol.24, November 2010.