An Empirical Study Of The Application Of Data Mining Techniques Classification In Intrusion Detection System

Mohssine EL AJJOURI, Siham BENHADOU, Hicham MEDROMI
Architecture System Team, Laboratory of Research in Engineering (LRI)
Hassan II University, ENSEM
Casablanca, Morocco

Abstract— To detect the unwanted traffic on a network, there is a great need for intrusion detection system robust and powerful which can detect the attacks. The IDS need to be accurate, adaptive, and extensible. Recently, many novel methods are experimented to build strong IDSs. These approaches include neural networks, decision trees, support vector machines, classification and clustering techniques, thereby helping in development of smart intrusion detection systems. The aim of this paper is to present an experimental study of many techniques of dataminig. The simulations are made using WEKA datamining tool and NSL-KDD dataset.

Index Terms— Intrusion Detection System, Data Mining, Attacks, Classification, NSL-KDD, WEKA .

1 Introduction

AN IDS analyzes and gathers information from differents field through a network or computer to determine potential security violations , which include intrusions attacks from abroad the company and misuse attacks from inside the company , this security management system is applied for computers and networks. For most organizations,since the severity of attacks occurring in the network has increased drastically, intrusion detection system have become a necessary addition to security infrastructure. Main causes of intrusions: accessing the systems by Attackers, privileges are misused by Authorized users , Authorized users of the systems who attempt to gain additional rights without autorisation.

IDS have two different approaches to detect intrusions [1]: misuse detection and anomaly detection. In misuse detection, attacks based on the patterns extracted from known intrusions are discovered. In anomaly detection, some parametres ard difined by the system administrator like the baseline, state of the network traffic load, typical packet size protocol and breakdown. The network segments are monitored by the anomaly detector, their state is compared with the normal baseline, so the anomaly detector look for the anomalies.

- *Mohssine EL AJJOURI is currently pursuing Ph.D degree in computer sciences at ENSEM School, Hassan II University, Morocco.E-mail: e.mohssine@gmail.com.*
- *Siham BENHADOU is currently working as Associate Professor in department of computer sciences at ENSEM School, Hassan II University, Morocco. E-mail: siham.benhadou@gmail.com.*
- *Hicham MEDROMI is currently working as Research Director And a Full Professor in department of computer sciences at ENSEM School, Hassan II University, Morocco. E-mail: hmedromi@yahoo.fr*

Application of Data Mining techniques to intrusion detection systems has great interest recently, so many people associated Knowledge Discovery in Database (KDD) to data mining , however DM can be viewed as a single step towards. Knowledge discovery is a current version of KDD data set is proposed known as the NSL-KDD due to supposed basic problems of the KDD CUP'99 [2]. Benefit of Data mining lies in the fact that it can withdraw the unknown and needed knowledge, and regularities from host log data and the massive network data. Using data mining algorithmes have been largely used in the last two decades for intrusion detection, such as decision tree, naive bayesian, support vector machine, neural network, k-nearest neighbors, fuzzy logic model.

The organization of this article is done as follows: in Section 2, an overview of related works is presented, Section 3 describes a data mining classification technique, Section 4 is dedicated on the performance metrics, experimental setup is presented in Section 5 is dedicated on results and discussions and Section 6 concludes the carried out research and possible future works.

2 Related Works

Sahilpreet Singh et Meenakshi Bansal [3] presents four different algorithms: Voted Perception Multilayer Perception, Radial Base Function, and Logistic Regression. To check the performance, a series of experiments were conducted in WEKA data mining, all these neural based algorithms are implemented on this tool. In [4], the authors proposed K-means an unsupervised clustering algorithm with information gain for reduction and feature selection to construct a network intrusion de-

tection system, The NSL-KDD dataset [5] has been used in two methods using all the dataset features firstly and then in a reduced form (with the same clustering algorithm). Only 23 features are selected from the 41 features in the reduced form, Information Gain of the attributes is used, the results show that there is a significant decrease in learning time of the algorithm and an increase in the accuracy. Neethu B [6], proposed Network IDS framework that is a combination of Naïve Bayes and Principal Component Analysis algorithm. This approach shows that Time consuming is less, also detection rate is high and low cost factor is low. D. Mohit , et al [7] have proposed IDS based neural network to detect and classify attacks in various category, the system that contain the algorithm is splited into some modules : processing of packet , feature extraction , classifier, training, accumulation of packet, KDD dataset, decision module. It provide 94% of detection rate efficiency , also it give a classification of attacks in 10 categories. Sampat, et al [8] have developed an application of applying data mining techniques to construct intrusion detection models by using metaclassifier and fuzzy clustering technique, the method improves the performance results and impressive detection accuracy and detection rate.

3 Data Mining Classification Method

From differents angles, we can analyzing data and encapsulating it into helpful information, that can be used to increase revenue, cuts costs, or together , this process is called data mining,. Data mining practice is one of many analytical tools for examining data, so the relationships identified is summarized and categorized. Descovering correlations or patterns among dozens of fields in large relational databases , is technically the process of data mining, so in data mining applications, to examine data , a selection of parameters in input are used. For intrusion detection, at present, data mining process mostly has four basic modeles: classification, sequence, association, and clustering. Data mining is an advanced technology, It can manage large quantity of data information no need of the users subjective evaluation. In this section , an overview of some method of classification used in dataminig is presented.

3.1 Zero R

ZeroR is the elementary classification method which ignores all predictors and relies on the target. The majority of categories (class) is predicted simply by the ZeroR. A baseline performance is helpful to determined as a reference for other classification methods because no predictability power used in ZeroR.

3.2 Naïve Bayesian Classifier

Naïve Bayes algorithm is a strongly easier Bayesian probability, it's based on independence assumption [9]. This classifier is anomaly based. For its operation, it must be ensured that the characteristics have different probabilities of happened in attacks and in normal TCP traffic. In the research world, this classifier is largely used, very simple to implement it, parameters are easy to estimate, its accuracy is practical good in comparison to the other approaches and learning is fast even on very large databases, this classifier is an instance of linear classifier. Naive Bayes algorithmes are efficient classification tools that are simple to use and interpret. It's particularly suitable when the measurement of the independent space (i.e., number of input variables) is high (a problem known as the trouble of dimensionality).

3.3 Nearest Neighbors Classifier (KNN)

To classify a new instance, KNN is a method non-parametric used for regression and classification ,the distance of its k neighbors is checked from the training set to classify it, It's used to search the measure of euclidean distance . The nearest training instance to the test instance predicts the same class as this training instance [10]. The target data is regressively compared with a set of predefined associations/rules/ sequences using data mining technique. The normal behaviour of the network is compared with the target data, a set of trained labeled data incorporating information regarding potential attacks or malicious data that are harmful to the system.

3.4 Nearest Neighbors Classifier (KNN)

J48 is an algorithm developed by Quinlan Ross and used to generate a decision tree [13], It's a continuation of Quinlan's previous ID3 algorithm (Iterative Dichotomiser 3). C4.5 is generally known as a mathematical classifier because decision trees produced by this algorithm can be used for classification. The J48 decision tree classifier algorithm splits recursively a data set in harmony to tests on attribute values in order to separate the possible predictions.

4 Experimental Setup

In this section we elaborate our experimental setup consisting of the data set NSL-KDD, for this comparative study, WEKA environment is selected.

4.1 Nsl Kdd Dataset

NSL KDD Dataset is used to clarify some of the basic complication of the KDD'99 data set. Using this datasets significant data can easily shared with other researchers, Many types of analysis have been carried out by many researchers on the NSL-KDD dataset employing many techniques and tools with a unique objective to develop an effecient intrusion detection system.NSL-KDD includes 42 characteristics which are organized in the following 4 basic categories: Content Features, Traffic Features, Time-based Traffic Features, Hostbased Traffic Features. also the attacks are separated in four categories namely probe, u2r, dos and r2l. Table 1 shows that in the dataset, every instance has 42 attributes or features including target class.

TABLE 1

FEAUTURES OF NSL KDD DATASET

No	Feature Name	No	Feature Name
1	Duration	22	s_guest_login
2	Protocol_type	23	Count
3	Service	24	Srv_count
4	Flag	25	Serror_rate
5	Src_bytes	26	Srv_serror_rate
6	Dst_bytes	27	Rerror_rate
7	Land	28	Srv_rerror_rate
8	Wrong_fragment	29	Same_srv_rate
9	Urgent	30	Diff_srv_rate
10	Hot	31	Srv_diff_host_rate
11	Num_failed_logins	32	Dst_host_count
12	Logged_in	33	Dst_host_srv_count
13	Num_compromised	34	Dst_host_same_srv_rate
14	Root_shell	35	Dst_host_diff_srv_rate
15	Su_attempted	36	Dst_host_same_src_port_rate
16	Num_root	37	Dst_host_srv_diff_host_rate
17	Num_file_creations	38	Dst_host_serror_rate
18	Num_shells	39	Dst_host_srv_serror_rate
19	Num_access_files	40	Dst_host_rerror_rate
20	Num_outbound_cmds	41	Dst_host_srv_rerror_rate
21	s_host_login	42	Normal or Attack

4.2 Weka Environment

WEKA [11] (Waikato Environment for Knowledges Analysis) is a sequences of machine learning concept for techniques of data mining, so algorithms can be called from Java code or applied to dataset directly, it incorporate tools for association rules, classification, visualization, clustering , pre-processing data and regression. It is also adapted for developing new machine learning algorithms developements. WEKA is free and based on JAVA environment, open-source, non-commercial and under the GNU General Public License policy (Fig.1).

The following steps are carried out in WEKA:

- ✓ Select the dataset.
- ✓ Run the four classifier algorithms
- ✓ Compare the four classifiers.

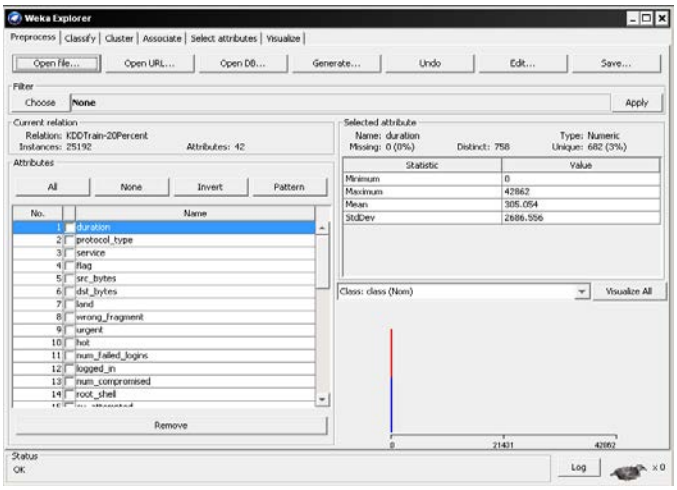

Fig 1. Weka Environment

4.2 Classfication Algorithm Comparison

This section explaines construction and evaluation of classification model. After launching ours experimentations, Table 2 shows the performance comparisons of classifiers based on traditional model accuracy in percentage, time taken to construct the classifiers in seconds, the totale of instances correctly classified and totale of instances incorrectly classified. For the model evaluation, the accuracy rate is the most used empirical measure but it's not sufficient enough.

TABLE 2

PERFORMANCE COMPARISONS OF CLASSIFIERS

Classifiers	Time	Efficiency	Correctly Classified Instances	Incorrectly Classified Instances
Zero R	0,44 s	53.386 %	13449	11743
NB	1,14 s	89.5919 %	22570	2622
KNN	0,05 s	86.4403 %	20051	5141
J48	13,38 s	87.5594 %	21081	4111

4.3 Evaluation Criteria

Intrusion detection is the procedure of analyzing events or monitoring that occur in a networked computer system or computer in order to detect behavior of users that conflict with the intended use of the system. When referring to the performance of IDSs, the following terms are often used when discussing their capabilities.

4.3.1 Accuracy

After classifying NSL KDD test dataset it's clearly shown that classifier Naïve Bayes shows the highest detection accuracy among all other algorithms. The graphical representation of evaluated parameter result is presented below :

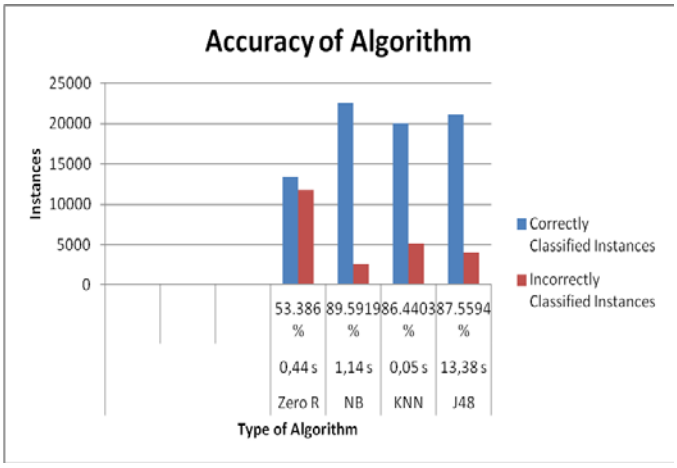

Fig 2. Accuracy of Algorithm

4.3.2 Kappa

The measure of agreement between categorical variables A and B is called Kappa, it's calculated by taking the accord expected by chance away from the observed accord and dividing by maximum possible accord, when the value is greater on zero, classifier is doing better than chance [12]. Figure 3 shows clearly after the validation step that Naive Bayes algorithm classifier performs accurate results than other algorithms. Classifier accuracy of Naive Bayes is 79% which is superior on comparison with the other algorithms.

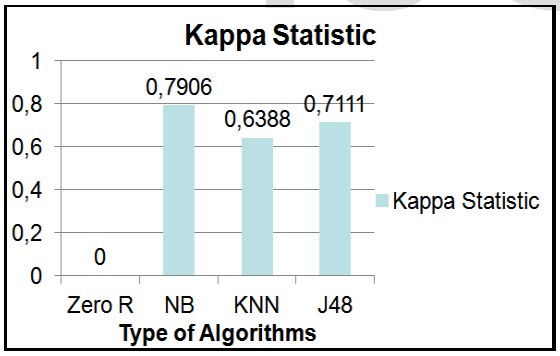

Fig 3. Kappa Statistic

4.3.3 Mean Absolute Error And Root Mean Squared Error

This criteria (MAE) determines the average magnitude of the mistakes, Fig 4 displays that the highest MAE rate is showed by Zero R algorithm, like mean absolute error, the RMSE is a variance at intervals of values predicted of a model and values observed of the modeled environment, From figure above, as compare to other algorithms, it's obvious that Zero R has the biggest RMSE rate.

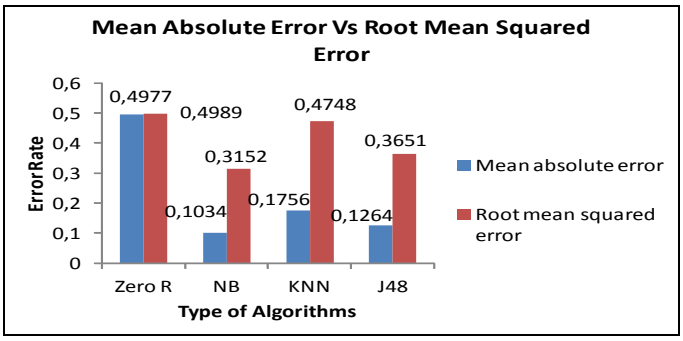

Fig 3. MAE Vs RMSE

4.3.4 Relative Absolute Error And Root Relative Squared Error

The relative absolute error is very resembling to the relative squared error in the meaning that it's also relative to a simple predictor, which is just the average of the values actual. Fig 5 clearly exposed that Zero R has the highest relative absolute error rate. RMSE is exploited to calculate variance at intervals of value observed and the value predicted. After doing cross validation, we assume that Voted Zero R evaluated the highest root relative squared error rate.

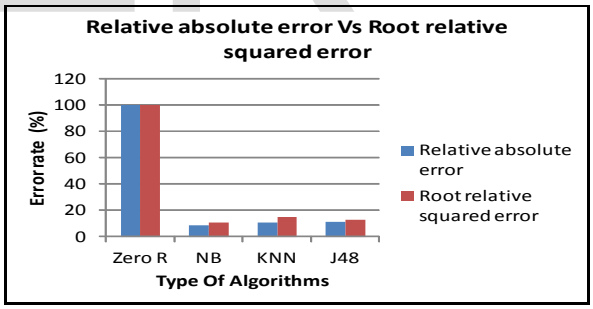

Fig 4. RAE Vs RRSE

4.3.5 Average Computational Time

Figure 6 indicate the average computational time taken by differents algorithms of data mining to test a model on training data. After classifying all algorithms on cross validation, it's clearly shown that the J48 has taken more time to build model on data training than the rest of among the algorithms of neural networks.

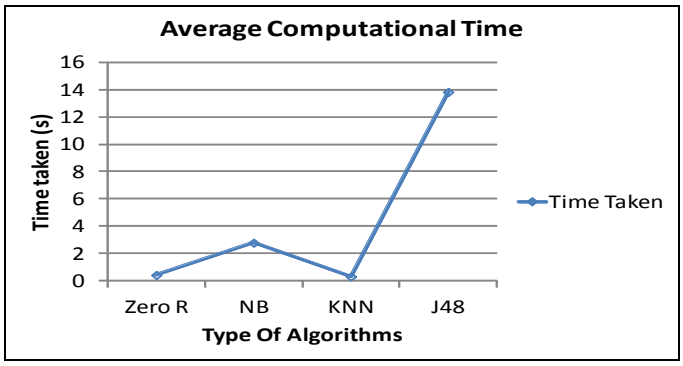

Fig 5. Computational time taken

5 Resuls And Discussion

We have performed experiments to evaluate the performance of classifiers like ZeroR, NB, KNN and J48 on our chosen NSL-KDD Dataset. These classifiers have been evaluated in WEKA environment using 42 attributes. It's obvious the naïve bayes performs betters by showing accuracy of 89,59%. In the case of Kappa, MAE, RAE also naïve bayes is the better with lowest error rate. The principal contribution of this paper is the result that prove naïve bayes algorithm is indeed better for classifying the Intrusion detection dataset.

6 Conclusion

The purpose of this experimental work is to judge the performance of classification algorithms of data mining. We use then weka to have a best performance comparison among the main popular classifier algorithms. After doing experimental work, it's clear that Naïve bayes has highest classification accuracy and lowest error rate as compared toother classifier algorithm. We showed also that datamining is an effective methodology which can be exploited on the field of intrusion detection, technique could be exploited also as a robust base in intrusion detection architecture for detecting new and unknown types of attacks. This result can be exploited as a base for developing our architecture of intrusion detection.

Acknowledgment

I would like to thank my advisor Pr. Siham BENHADOU and Pr Hicham Medromi, for their invaluable guidance and many useful suggestions during my work on this paper.

References

[1] M. El ajjouri, S. Benhadou, and H. Medromi , "New Collaborative Intrusion Detection Architecture Based on Multi Agent Systems," Journal of communication and computer, vol. 13 , pp. 1-10, January 2016.

[2] Amneh H. Alamleh , "Network Intrusion Classification Using Data Mining Techniques", Thesis, Faculty of Graduate Studies Zarqa University – Jordan , August, 2015.

[3] S. Singh, M. Bansal , "Improvement of Intrusion Detection System in Data Mining using Neural Network," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3 , issue 9, Septembre 2013.

[4] D. Yousif Mahmood, M. Abdullah Hussein , "Feature Based Unsupervised Intrusion Detection," International Journal of Computer, Electrical, Automation, Control and Information Engineering , vol. 8, N 9, 2014.

[5] The Knowledge Discovery in Databases, NSL-KDD dataset, http://nsl.cs.unb.ca/NSL-KDD/

[6] B. Neethu, "Classification of Intrusion Detection Dataset using machine learning Approaches," International Journal of Electronics and Computer Science Engineering , 1044, ISSN- 2277-1956.

[7] D. Mohit, K. Gayatri, M. Vrushali , G. Archana , B.Namrata , "Using Artificial Neural Network Classification and Invention of Intrusion in Network Intrusion Detection System," International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, Issue 2, February 2015.

[8] R. Sampat, S. Sonawani, "Network Intrusion Detection Using Dynamic Fuzzy C Means Clustering,", International Journal of New Technologies in Science and Engineering, Vol. 2, Issue. 4, October 2015.

[9] S. Mukherjeea, N. Sharmaa, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction," Procedia Technology 4 (2012) 119 – 128.

[10] M. Pandey, S. Taruna, "A Multi-Level Classification Model Pertaining To The Stuednts's Academic Performance Prediction," International Journal of Advances in Engineering & Technology, Sept., 2014.

[11] http://www.cs.waikato.ac.nz/ml/weka/

[12] S. Singh, M. Bansal, "Improvement of Intrusion Detection System in Data Mining using Neural Network," International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, Issue 9, Septembre 2014.

[13] Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993).

[14] M. Mohammada, N. Sulaimana, O. Muhsinb, "A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment," Procedia Computer Science 3 (2011) 1237–1242.