A Data Mining Approach for Attribute Selection in Intrusion Detection System

Richa Pandey
M. Tech Scholar
School of Computing
Graphic Era Hill
University Bhimtal
Campus Uttarakhand,
India

Janmejay Pant
(CSE)
School of Computing
University Bhimtal
Campus Uttarakhand,
India

ABSTRACT
Data Mining is a collection of tools and techniques for extraction of useful data from large amount of databases and now a days there are many intruders who try to steal useful data or to change the originality of data.An Intrusion Detection System (IDS) is a method which is use for defence method which check the activities of the computer network and reports the malicious activities to the network administrator if there is any. As the Intruders do more than one attempts to gain access to the network and try to destroy the authentication of the organization's data. The security which is the main issue for any organization have to take steps for maintaining the originality of the data.

Thus intrusion detection field has been an important research issue in today's world. In this paper we are going to discover an approach for attribute selection which helps in improvement of the accuracy which will be shown by ROC curve which is Receiving Operating Characteristic Curve. By gaing the results by do algorithm we will have the best way to improve in the curve.

Keywords
IDS, WEKA, ROC curve, KDD process

1. INTRODUCTION
Data Mining is method of extracting useful data from huge sets of data. The other way of defining it is that data mining is mining knowledge from data.

There is large amount of data available in the Information sector and business sector and data is of no use until it is converted into some useful information. Data analysis is necessary for huge amount of data and extract useful information from data.

An intrusion detection system (IDS) is a device or software application that monitors a network or systems for malicious activity or policy violations[1]. When any activity is detected it is reported either to an administrator or collected centrally using a security information and event management (SIEM) system. A SIEM system combines outputs from multiple sources, and uses alarm filtering techniques to distinguish malicious activity from false alarms[1].

2. PROBLEM FORMULATION
The problem faced during the analysis of algorithm named naïve bayse that ROC curve was very poor so we decide to analyze J48 algorithm with attribute selection.We have use WEKA tool for analysis.

3. DATA MINING STEPS
The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages:

(1) Selection
(2) Pre-processing
(3) Transformation
(4) Data Mining

Interpretation/Evaluation

4. FEATURE SELECTION AND ATTRIBUTE SELECTION
The large amount of data transfer over the network and real time intrusion detection is near to impossible. Feature selection help reduce the computation time and complexity of model. Feature selection is the matter of concerns in 60s also.

5. KDD
The KDD CUP 1999 benchmark datasets are used in order to evaluate different feature selection method for Intrusion detection system. It consists of 4,940,000 connection records.

Each connection had a label of either normal or the attack type, with exactly one specific attack type falls into one of the four attacks categories as: Denial of Service Attack (DoS), User to Root Attack (U2R), Remote to Local Attack (R2L) and Probing Attack.[2]

- **Denial of Service Attack (DOS)**: Attacks of this type deprive the host or legitimate user from using the service or resources.
- **Probe Attack**: These attacks automatically scan a network of computers or a DNS server to find valid IP addresses.
- **Remote to Local (R2L) Attack**: In this type of attack an attacker who does not have an account on a victim machine gains local access to the machine and modifies the data.
- **User to Root (U2R) Attack**: In this type of attack a local user on a machine is able to obtain privileges normally reserved for the super (root) users.

When we detect network intrusion mainly four situations are arise: True negative, True Positive, False Negative and False Positive for actual and predicate class in confusion matrix. [3]

- True Negative: Stand for number of detected normal example and these are normal in actual.
- True Positive: Stand for number of detected attacks class and these are an anomaly actually.
- False Negative: Stand for number of detected normal example but they are anomaly in actual.
- False Positive: Stand for number of detected attack class but in actual they are belongs to normal class.

6. SIMULATION AND RESULT

The tool which I have use for simulation purpose is WEKA. For simulation purposes and analysis we use WEKA tool. Waikato Environment for

Knowledge Analysis (WEKA) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License.

WEKA (pronounced to rhyme with Mecca) is a workbench that contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The original non-Java version of WEKA was a Tcl /Tk front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data pre-processing utilities in C, and a Make file-based system for running machine learning experiments. Advantages of WEKA include:

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces. [4]

6.1 Steps Included

1. Firstly we will conclude the drawback of Naïve Bayes algorithm with respect of ROC curve.
2. Now we will analyse J48 algorithm with respect to the ROC curve.
3. Evaluate the curve
4. Obtain results.

Now following graphs will show the evaluation of results:

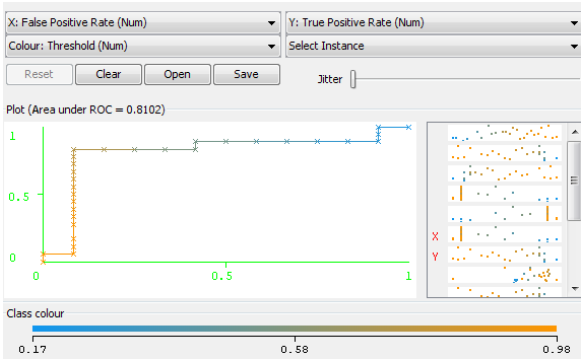

Fig 1:Naïve Bayes C1 ROC curve.

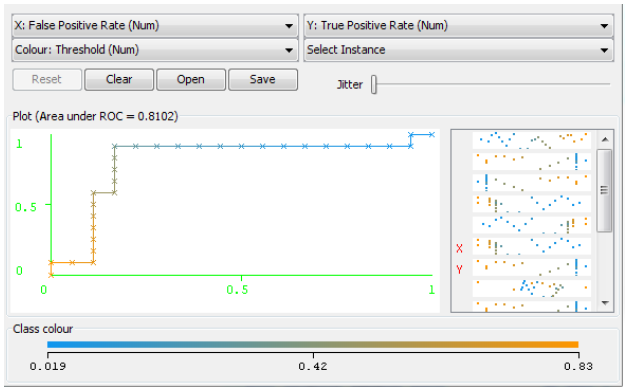

Fig 2:Naïve Bayes C1 ROC curve.

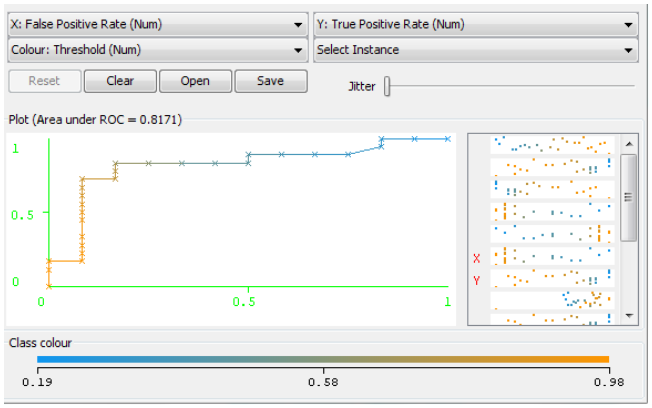

Fig 3: Naïve bayes after attribute selection C 0 ROC curves.

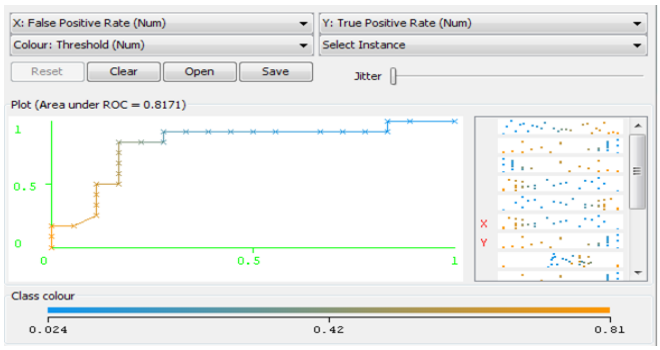

Fig 4: Naïve bayes after attribute selection C 1 ROC curves.

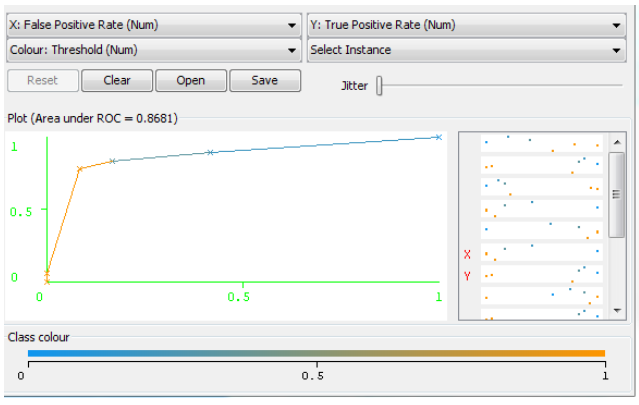

Fig 5: J48 C 1 ROC curve.

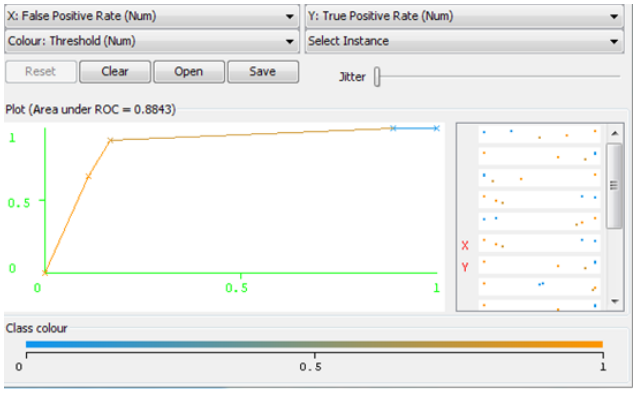

Fig 6:J48 C 0 ROC curve.

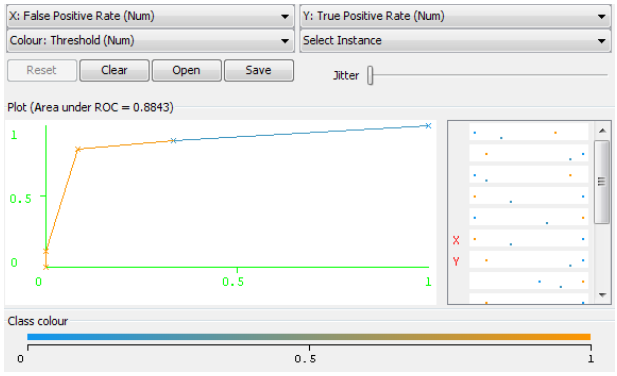

Fig 7:J48 C 1 ROC curve.

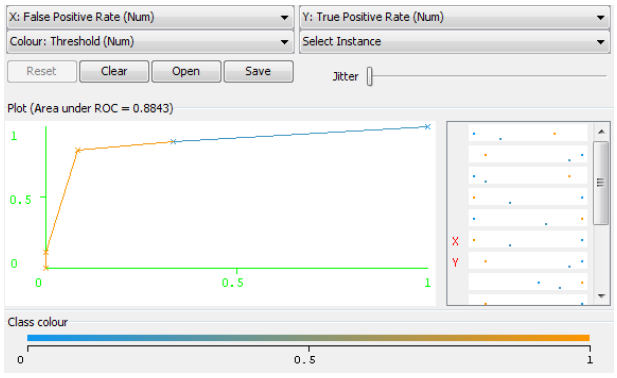

Fig 8:J48 C 0 ROC curve.

6.2 ROC curve measures:

AUC	Quality of Test
0.9-1	Excellent
0.8-0.9	Good
0.7-0.8	Fair
0.6-0.7	Poor
0.5-0.6	Fail

Since from above analysis we conclude the following results:

1.Analysis from the Naïve Bayes algorithm

- When we take whole dataset of KDD for IDS the ROC curve value is 0.8102.
- When we select some attribute the value is slightly improved i.e. 0.8241.

2.Analysis KDD Data set for IDS using J48 algorithm

- When whole dataset is taken from KDD for IDS the ROC curve value is 0.8681
- When whole dataset from KDD is used for IDS the ROC curve value is 0.8843

7. CONCLUSION

The following conclusion comes out of work**:**

1. Analysis of two algorithms.
2. J48 is better algorithm than Naïve bayes
3. Naive bayes works better after attribute selection.
4. J48 works betters after attribute selection.

8. REFERENCES

[1] Intrusion detection system https://en.wikipedia.org/wiki/Intrusion_detection_system

[2] International Journal of Advances in Engineering & Technology, July 2013. ©IJAET ISSN: 22311963 1319 Vol. 6, Issue 3, pp. 1319-1324 Feature selection using random forest in intrusion detection system Sneh Lata Pundir and Amrita Department of CSE, Sharda University, Greater Noida, India

[3] Ashalata Panigrahi and Manas Ranjan Patra, "An ANN Approach for Network intrusion detection using entropy based feature selection". International Journal of Network Security & Its Applications (IJNSA), Vol.7, No.3, May 2015

[4] An Introduction to the WEKA Data Mining System by Zdravko, Ingrid Russell University of Hartford.

[5] Nsl-Kdd data set for network-based intrusion detection system. Available: http:// iscx.ca/NSL-KDD/.

[6] Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution Lei Yu Huan Liu

[7] Shelly Gupta et al./ Indian Journal of Computer Science and Engineering (IJCSE) DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS

[8] Journal of Machine Learning Research 11 (2010) 2533-2541 Submitted 6/10; Revised 8/10; Published 9/10 WEKA—Experiences with a Java Open-Source Project

[9] Harshit Saxena, Vineet Richaariya, "Intrusion Detection in KDD99 Dataset using SVM-PSO and Feature Reduction with Information Gain", International Journal of Computer Applications (0975 – 8887) Volume 98– No.6, July 2014.

[10] Nupur N. Majethiya and Dipak C. Patel, "Efficient Intrusion Detection System with Reduced Dimensionality", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 4, Issue 2, March-April 2015 ISSN 2278-6856.

[11] Abebe Tesfahun, D. Lalitha Bhaskari "Effective Hybrid Intrusion Detection System: A Layered Approach", I. J. Computer Network and Information Security, 2015, 3, 35-41.

[12] Sunil Choudhary and Pankaj Dalal, "An Architecture for Network Intrusion Detection System based on DAG Classification", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 4, April 2015.

[13] Venkata Suneetha Takkellapati et.al "Network Intrusion Detection system based on Feature Selection and Triangle area Support Vector Machine" International Journal of Engineering Trends and Technology- Volume 3 Issue 4- 2012.