

IDENTIFICATION AND CLASSIFICATION OF INDIAN CYBER CRIME USING ENHANCED NAIVE BAYES ALGORITHM

***Rajbir Kaur, Er. Roop Lal and Prof. Gurpreet Singh**

¹M.Tech CSE St. Soldier Institute of Engg. & Technology, Near NIT, Jalandhar.

²Assistant Professor, Department of Computer Sc. & Engg. St. Soldier Institute of Engg. & Technology, Near NIT, Jalandhar.

³Professor, Department of Computer Sc. & Engg. St. Soldier Institute of Engg. & Technology, Near NIT, Jalandhar.

Article Received on 25/02/2021

Article Revised on 17/03/2021

Article Accepted on 07/04/2021

***Corresponding Author**

Rajbir Kaur

M.Tech CSE St. Soldier
Institute of Engg. &
Technology, Near NIT,
Jalandhar.

ABSTRACT

The current day, changes in public activity style and conditions of living make the people to run over marvels called cybercrime. Different organizations, for example, POLICE Department, Cyber Cell, CBI are working thoroughly to battle the wrongdoing. In any case, the difficulties to dissect the wrongdoing and capture the crimes

is getting more troublesome as the crime percentage is expanding numerous Algorithm have been projected by the scientists for successful examination. The primary drawback is that the volume of information as for the wrongdoing exercises and hoodlums expanded and there is an incredible requirement for breaking down the information, henceforth to have a superior model the information about the wrongdoing and the criminal consistently is consistently beneficial. This idea has driven towards the utilization of information digging strategies for breaking down this voluminous information.

I. INTRODUCTION

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny,

Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

II. ENTROPY

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.

III. INFORMATION GAIN

The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches).

IV. NAIVE BAYES ALGORITHM

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

V. TYPES OF CYBERCRIME

In order to protect yourself you need to know about the different ways in which your computer can be compromised and your privacy infringed. In this section, we discuss a few common tools and techniques employed by the cyber criminals. This isn't an exhaustive list by any means, but will give you a comprehensive idea of the loopholes in networks and security systems, which can be exploited by attackers, and also their possible motives for doing so.

HACKING

VIRUS DISSEMINATION

LOGIC BOMBS

DENIAL-OF-SERVICE ATTACK
EMAIL BOMBING AND SPAMMING
WEB JACKING
CYBER STALKING
DATA DIDDLING
IDENTITY THEFT AND CREDIT CARD FRAUD
SALAMI SLICING ATTACK
SOFTWARE PIRACY

VI. OBJECTIVES OF THE STUDY

1. To identify the cybercrime criminal activities with Enhanced Naive Bayes Algorithm
2. To analyzes the vulnerabilities of each technique and provides recommendations for the development of a cybercrime detection model that can detect cybercrimes effectively compared with the existing techniques.
3. To compare the Classification Accuracy based on correctly classified instances, incorrect classified instances and error rate.

VII. Tools Used for Research Work

Netbeans Ide

It is an integrated development environment for developing primarily with java but also with other languages. It is also an application platform framework for java desktop applications and others. The Net Beans Platform allows applications to be developed from a set of modular Software components called modules. Net Beans IDE supports all Java application types. All the functions of the IDE are provided by modules. Each module provides a well-defined function, such as support for the Java language, editing, or support for the CVS versioning system.

WEKA

WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in java. WEKA is free software available under the General Public License. The WEKA (pronounced Weh-Kuh) workbench. It contains a collection of visualization tools and algorithms for data analysis and predictive modeling together with graphical user interfaces for easy access to this functionality.^[40] The front-end to (mostly third-party) modeling algorithms implemented in other programming languages.

RESULTS

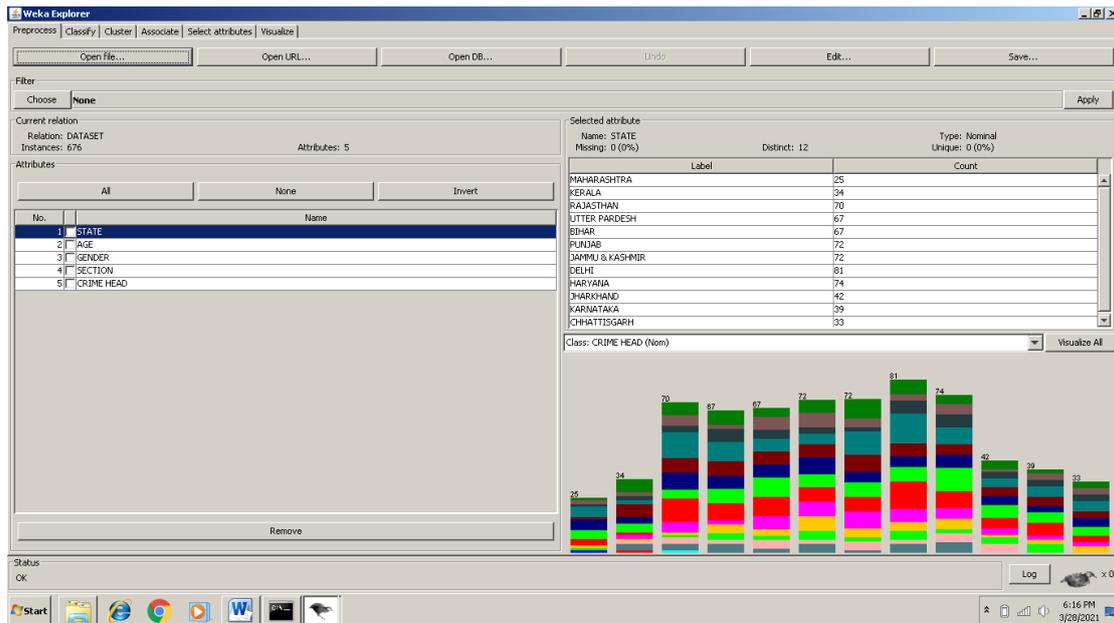


Fig. 1: Describes the “State” first attribute which is an anominal attribute. In case of nominal attribute WEKA specifies the various values (labels) under that attribute along with number of instances under that label of the attribute. Graph shows the distribution of “Crime Head” class attribute in “State” attribute.

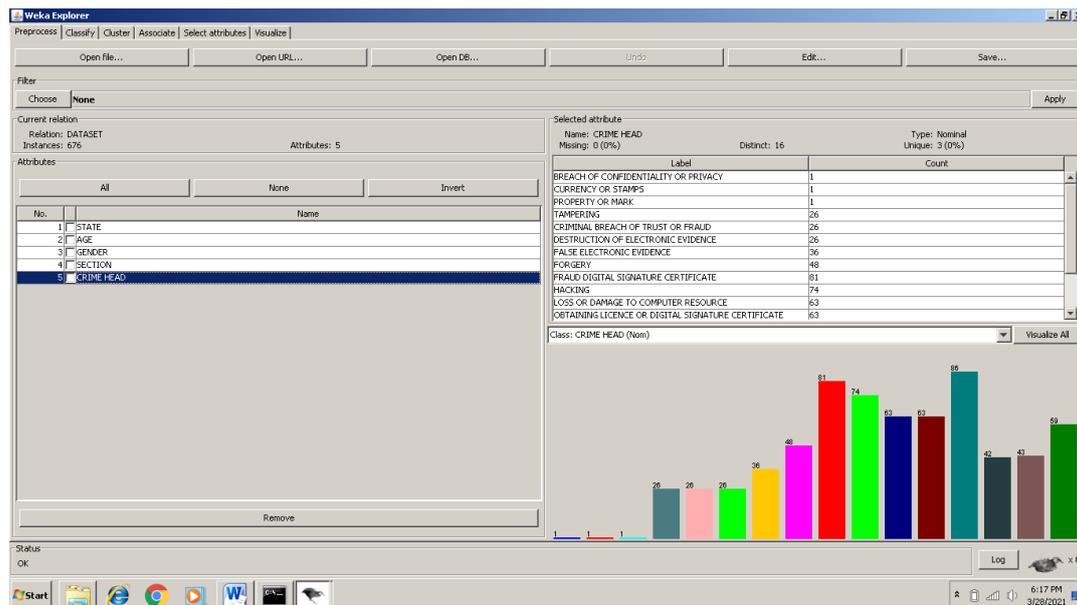


Fig. 2: Shows the main WEKA Explorer interface with the dataset loaded. The last attribute “Crime Head” is taken as a class attribute by the WEKA. This attribute contains Sixteen states. The count of number of instances under each service in the dataset is shown numerically as well as graphically.

Table 1: Comparison between Random Forest and ENBA on the basis of Correctly Classified Instances, Incorrectly Classified Instances and Error Rate.

Parameters	Random Forest	ENBA
Correctly classified instances	12.72	24.7
Incorrectly Classified Instances	87.27	75.29
Error rate	99.99	93.10

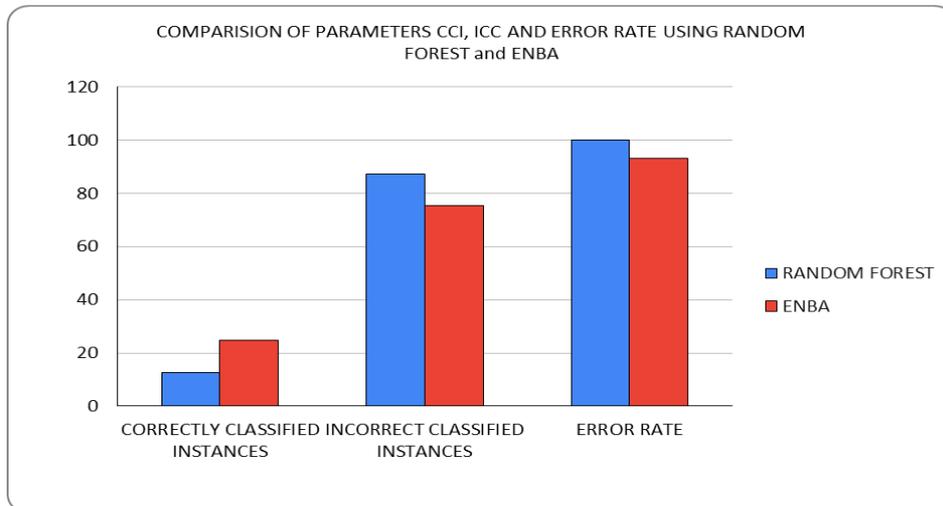


Fig. 3: Chart shows the comparison between Random Forest and ENBA on the basis of Correctly Classified Instances, Incorrectly Classified Instances and Error Rate.

The Correctly Classified Instances, Incorrectly Classified Instances and Error Rate esteems for Enhanced Naïve Bayes Algorithm is superior to other traditional Algorithm.

CONCLUSION

The research work is to develop a simple approach to detect the risk of cybercrime activities. We have used dataset of previously cybercrime activities in india and applied data mining classification techniques on it. The knowledge acquired by the usage of data mining techniques can be used to make successful and effective decisions that will improve and restrict the future cybercrime activities. Data set contains of 676 instance and five attributes. In this research we presented the systematic method of crime detection at national level. As crime data is increasing now to control the crime is again become a difficult task so to solve this problem author is presenting a systematic way of mining crime data detection classification in such a way so that it's become easy to solve the crime problem throughout the world.

Two algorithms are used under weka and the comparisons are made based on the accuracy among these classifiers and different error measures are used to determine the best classifier.

Experiments results show that Enhanced Naïve Bayes Algorithm has the best performance among other classifier. The comparison table clearly declare which classification model is better The two algorithm are works well but here Enhanced Naïve Bayes Algorithm works more accurate than other algorithm. In future work, more dataset instance will be collected and will be compared and analyzed with other data mining techniques such as association and clustering.

REFERENCES

1. Dr. Neelam Sahu, Sagar Darokar, “Analysis Of Cyber Crime Data Using Data Mining Technique I E S R J, FEB 2018; 4(2).
2. Sankara Moorthy, Densingh Joshua Israel, “Analysis of Indian cybercrime dataset for age demography”, International Journal of Applied Engineering Research, January 2015.
3. Shiza Andleeb, Dr. Rashid Ahmed, Zaheer Ahmed, Maira Kanwal, “Identification and Classification of Cybercrimes using Text Mining Technique”, International Conference on Frontiers of Information Technology (FIT), IEEE, 2019.
4. Twinkle Arora, 2Monika Sharma, 3Sunil Kumar Khatri, “Detection of Cyber Crime on Social Media using Random Forest Algorithm” 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC), IEEE, 2019.
5. Kim-Kwang Raymond Choo, Mauro Conti, “Special Issue on Big Data Applications in Cyber Security and Threat Intelligence – Part 1”, IEEE TRANSACTIONS ON BIG DATA, JULY-SEPTEMBER, 2019; 5(3).
6. Ildiko Pete, Yi Ting Chua, “An Assessment of the Usability of Cybercrime Datasets”, IEEE, 2019.
7. Priyanka Dhaka, Rahul Johari, “CRIB: Cyber Crime Investigation, Data Archival and Analysis using Big Data Tool”, International Conference on Computing, Communication and Automation (ICCCA2016), IEEE, 2016.
8. Mohiuddin Ali Khan¹, Sateesh Kumar Pradhan², Huda Fatima³, “Applying Data Mining Techniques in Cyber Crimes”, IEEE, 2017.