Performance Evaluation of Multi Navie Bayes Algorithm for Spam Email Detection

[1]Komal Ahuja, [2]Amandeep
M. Tech (CSE Department)
GRIMT Radaur, Haryana, India

***Abstract*— In this paper we are define how to detect new malicious executables or Malware by using Data mining (DM) namely multi-naive with the variable Byte sequence. It is very important to detect and delete new malware in an effective manner. In this paper, we propose three data mining (DM) algorithms to produce new classifiers with separate features: Bayes Net, Naïve – Bayes and a Multi Classifier system and the comparison between three classification methods. It comprises of root kit data collection, classification, data pre-processing and performance evaluation phases.**

***Keyword* - Data Mining, Classification, Security, Prediction, Multi-naïve Bayes**

I. INTRODUCTION

In now days a serious security threat is malicious executables, especially, new unseen malicious executables arriving as email attachments. These new malicious executables are build at the rate of thousands every year and suffering a serious security threat. Current anti-virus systems try to detect these new malicious programs with heuristics generated by hand. This technique is costly and effective. A malware executable is defined to be a program that shows a malicious function, i.e. compromising a system's security, obtaining sensitive information without the user's permission. Using data mining (DM) methods, their limitation is to automatically design and build a scanner that accurately detects malicious executables before they given a chance to run.

Their goal in the evaluation of this method was to simulate the task of detecting new malicious executables (malware). To do this we divided their data into two sets: a training set and a test set is used with standard cross-validation methodology. The training set was used by the data mining (DM) algorithms to create classifiers to classify previously unseen binaries as benign. A test set is collection of dataset that had no examples in it that were seen over the training of an algorithm. This subset was used to test an algorithms' performance over similar, unseen data and its performance in new malicious executables. Both the test and training data were malicious executables collect from public sources.

MALWARE:

Malicious software which enters in a system without authorization user of the system. The term is created from merging the words 'malicious' and 'software'. Today's Malware is a big threat in computing world. It continues to grow in volume and evolve in complexity. As more organizations try to address the problem, the number of websites distributing the malware is increasing at an alarming rate and is getting out of control. Many malware enters the system while downloading files over Internet. The malicious software finds its way into the system, it scans for vulnerabilities of operating system (OS) and perform unintended actions on the system and finally slowing down the performance of the system.[20]

TYPES OF MALWARE

In recent time, the number of information security threats caused by malware has rapidly increased, which leads to studying the threats and categorizing them, to simplify the process of discovering and handling them, in order to detect them and find solutions. Malware has been categorized into seventeen different types. In this section we have listed and discussed the most common categories as follows:

Figure 1:- Types of Malware

A. Virus

Virus is a computer program that has the ability to harm and self-replicating in order to infect host; viruses are linked or attached to a software utility (e.g. PDF document). Launching the infected PDF document could then prompt the virus, and a sequence of events may occur depend on the function of the virus.

B. Worm

Another kind of harmful programs is worm, which can replicate itself and invisibly transfer through networking. The effects of worms differ from viruses as the former need help from any file, to work and mainly its effect is on networking bandwidth or sending junk emails. One example of worms is Conficker.

C. Spyware

This may occur, when users download free or trial software. In this kind, the users are observed by spies; hence their passwords, account numbers and every other personal detail become vulnerable.

D. Adware

This kind usually happens, while downloading free games or it is combined and embedded with advertisements, so when we watch advertisements this embedded code is installed to our PCs. This kind aims to observe the user's activities, when using networking.

E. Trojan

This kind gives power to remote hijackers, to use your system as they wish. They may get your passwords, observe your systems or damage the system files.

F. Botnet

This type of malware controls your systems remotely and sends spam or spyware. Most of botnets are zombie and wait.[16]

1.2 Malware Obfuscation

The obfuscation means modifying the program code in a way to preserve its functionality with the aim to reduce vulnerability to any type of static analysis and to deter reverse engineering by making the code difficult to understand and less readable. Obfuscation techniques i.e. packing, polymorphism and metamorphism are used by malware authors legitimate software developers. They are use code obfuscation techniques for different reasons. Code obfuscation is effectively used by malware authors to avoid antivirus scanners since it modifies the program code to generate offspring copies which have the same functionality, but with different byte sequence and virus signature is not recognized by antivirus scanners.

1.3 Windows API Calls

Windows API calling reflects the behaviour of executables. The Windows API function calls fall under many functional levels i.e. Libraries, user interfaces, network resources, windows shell and system services. After all the API calls reflect the functional levels of a program, search of the API calls would lead to an understanding of the behaviour of the file. Malicious codes (malware) are able to disguise their behaviour by using API functions give under Win32 environment to implement tasks. Consequently, in binary static analysis, the focus is on examine all documented Windows API call features to understand the malware behaviour.

In the Windows operating system, user applications build on the interface give within a set of libraries, i.e. KERNEL32.DLL, USER32.DLL, and NTDLL.DLL in order to key system resources including files, network information, and the registry. This interface is called Win32 API. Applications call functions in NTDLL.DLL known as the Native API. The Native API functions operate system calls in order to have the kernel provide the requested service. The extracted calls are detain to those that affect the files. Many features related to the calls that generated files or even get information from the file to change some value.

MALWARE PROPAGATION

Many studies and researches focused on analyzing the malware propagation in the communications, digital world, s and computer networks, some of the pattern and experimental procedures have been to examine the effect of malware and the way it generate in these fields.

- **Through Operating System:** Malware is attacking the operating systems (OS) i.e. Mac, Windows, Android, and Linux, but not in the equal level and strength because some operating systems (OS) have more defense mechanisms which don't confess the malware to attain its design purpose.

- **Through Wireless Networks :** We have introduced the mobile and Smartphone applications and some security concern related to wireless networks., The Bluetooth technology has been introduced in particular project named as Blue Bag that contain a covert attack and scanning device, which demonstrates how attackers can affect and reach a high range of mobiles and devices running a Bluetooth Technology, they have find ***a*** some weaknesses in Bluetooth technology, which may allow attackers.

- **Through File Sharing:** File sharing has become a very common application for Peer-to-Peer networking, which allows the users to share a huge number of digitally stored information. The most common file sharing networks is Kazaa, which has been developed in 2001, depends on the Fast Track Protocol, Kazaa was subsequently under license as a legal music subscription service, but as of August 2012, the Kazaa website is not offering a music service anymore.

- **Through Social Networking:** During the last few years, online social networks have become very popular and grew tremendously as they act as platform of real-world relationship; its popularity comes from the feature of virtual interaction techniques.

II. RELATED STUDY

Suvendu Jena et.al (2015) Data mining (DM) is the process of posing queries and extracting patterns, often previously unknown from more quantities of data using pattern matching or many other reasoning techniques. Data mining (DM) has many applications in security including for national security as well as for cyber security. A serious security threat today is malicious executables, especially new, unseen malicious executables often arriving as email attachments. These new malicious executables are generating at the rate of thousands every year and create a serious security threat. [1]

Anirudh Harisinghaney et.al. (2014) Internet has changed the mode of communication, which has become more concentrated on emails. Emails, online messenger, text messages and chatting have become part or parcel of our lives. Out of all communications, emails are more prone to exploitation. Thus, many email providers employ algorithms to filter emails depends on spam and ham. Our aim is to detect text as well image based spam emails. To obtain the objective we applied three algorithms i.e.KNN algorithm, Nai·ve Bayes (NB) algorithm and reverse DB SCAN algorithm 2]

Somayeh Soltani1et.al. (April 2014) The author modify the destructive effect of botnets is a concern of security scholars. Though various mechanisms are proposed for real world botnets, botnets detection, still keep and do their destructive operations. Botnets have introduced by new evasion techniques and covert communication channels. The characteristics of real world botnets help security researchers in developing more powerful detection methods. There are some surveys in the literature study the botnet detection methods. yet they do not observe to real world botnets a lot. The author study many aspects of several real world botnets, such that Conficker, Kraken, Rustock, communication [3]

Rajkumar E.V.1 Aravindharamanan (2014) Securing the web is a huge challenge that the modern era of computers have seen. Day by day the threat levels increase day by day making the network accessible to attacks. Many new strategies are involved into the field of cyber security to protect websites from attacks. But however malware has remained a major cause of analyze to web developers and server administrators. [4]

Kirti Mathur et. al. (2013) The computer technology has arrive as a necessity in our day to day life to deal with several aspects like education, communication, banking, entertainment etc. Computer system security is threatened by weapons named as malware to arrive malicious intention of its writers. Several solutions are available to detect these threats like AV Scanners, Firewalls, Intrusion Detection System (IDS), and etc. The malware detection traditionally uses signatures of malware to detect their presence in our system. But these methods are also avoided due to some obfuscation techniques operating by malware authors. [5]

Anshul Goyal and Rajni Mehta (2012) with the broad application Classification is an important data mining (DM) technique. It classifies data of many kinds. Classification is used in every field of our life. Classification is used to analyze each item in a set of data into one of pretend set of classes and groups. This author has been carried out to build a performance evaluation of Naïve Bayes and j48 classification algorithm. Naive Bayes (NB) algorithm is depend upon probability and j48 algorithm is depend upon decision tree. [6]

Ammar Ahmed E.et.al. (2012) A malware is a program that has malicious intent. In nowadays, malware authors apply many sophisticated techniques i.e. packing and obfuscation to detect malware detection. That built zero-day attacks and false positives (FP) the most challenging problems in the malware detection field. The author study static and dynamic analysis techniques that are used in malware detection. Static analysis techniques, dynamic analysis techniques and combination including Signature-Based technique. [7]

Parisa Bahraminikoo (Sep.-Oct. 2012) malware is any software that gives full control of your computer to do whatever the malware creator wants. Malware can be a virus, Trojan, worm, root kit, etc. Spyware is a type of malicious software installed on computers that collects information about users without their knowledge. Artificial Intelligence was established at Dartmuth College during a conference. The technology developed so much that it started contains many other branches of engineering i.e. electronics, robotics etc. [8]

Jianyong Dai, et.al.(may 2009) The author present a novel approach to detect unknown virus by using dynamic instruction sequences (IDS) mining techniques. We gather runtime instruction sequences from unknown executables and analyze instruction sequences into basic blocks. The author assemble instruction sequence patterns depends on three types of instruction associations within derived basic blocks. Following a data mining process, The author perform feature extraction, feature selection and then make a classification model to learn instruction association patterns from benign and malicious dataset. [9]

Bhavani Thuraisingham et. al.(2008) discuss many data mining techniques that we have successfully applied for cyber security. These applications contain but are not limited to malicious code detection by mining binary executables, network intrusion detection by mining network traffic, data stream mining and anomaly detection. We summarize our achievements and current works at the University of Texas at Dallas on intrusion detection and cyber-security research. [10]

III. PROPOSED WORK

Multi-Naïve Bayes classifier is used for detection of new malicious executables i.e. email attachment. With the help of proposed classifier we can filter out the malicious email. it is very important to detect and delete new malware in an effective manner. It provides the security to email. This method is required because even using a machine with one gigabyte of RAM, the size of the binary data is too large to fit into computer's memory. The Naive Bayes algorithm requires a table of all strings or bytes to compute its probabilities.

To solve this problem we divided the problem into smaller pieces that would fit in memory and trained a Naive Bayes algorithm over each of the sub problems. We split the data evenly into several sets. For each set we trained a Naive Bayes classifier. Final prediction for a binary is the product of the predictions of the n classifiers. In our experiments we will use 8 classifiers (n = 8). More formally, the Multi-Naive Bayes promotes a vote of confidence between all of the underlying Naive Bayes classifiers. Each Naive Bayes classifier gives a probability of a class C given a set of bytes F which the Multi-Naive Bayes uses to generate a probability for class C given F over all the classifiers.

IV. EXPERIMENTAL RESULT AND ANALYSIS

We have used the open source software suite WEKA which stands for Waikato Environment for Knowledge learning. The main reason why I selected WEKA was because of its versatility. WEKA is used in many different application areas; it is used for educational purposes and research. WEKA tool used for data analysis, machine learning and predictive modelling WEKA, formally called Waikato Environment for Knowledge analysis is a computer program that was developed at the University of Waikato in New Zealand for the identifying information from raw data collected from agricultural domain.

WEKA supports many different data mining tasks i.e. data pre-processing, regression, classification, clustering, visualization and feature selection. Weka's techniques are predicated on the inference that the data is available as a single flat file , where every data point is described by a fixed number of attributes i.e. numeric or nominal attributes, but some other attribute types are also supported. The basic premise of the application is to handle a computer application that can be trained to machine learning capabilities and provide useful information in the form of trends and patterns. WEKA is an open source application that is freely available under the (GNU) general public license agreement. Originally, it is written in C the WEKA application has been completely rewritten in Java and is compatible with every computing platform. WEKA is user friendly with a graphical interface that allows for quick set up and operation.

Firstly. The proposed multi Naive Bayes algorithm is used in the weka. And check the detailed accuracy, TP Rate, FP Rate, ROC Area, MCC,F-Measure etc The following table and graph predict the analysis that has been made:

Table 1 Comparison between existing classifier and proposed classifier

Sr No	Parameter	NaiveBayesUpdateble	NaiveBayesMultinomialUpdateable
1.	TP Rate	0.793	0.801
2.	FP Rate	0.152	0.195
3.	Precision	0.842	0.808
4.	F-Measure	0.794	0.803
5.	MCC	0.632	0.596
6.	ROC Area	0.931	0.86
7.	PRC Area	0.922	0.847

This table contains results of all algorithms for some parameter like ROC Area, precision, recall and f-measure etc. From the results it is concluded that results of multi naïve bias is better than all others.

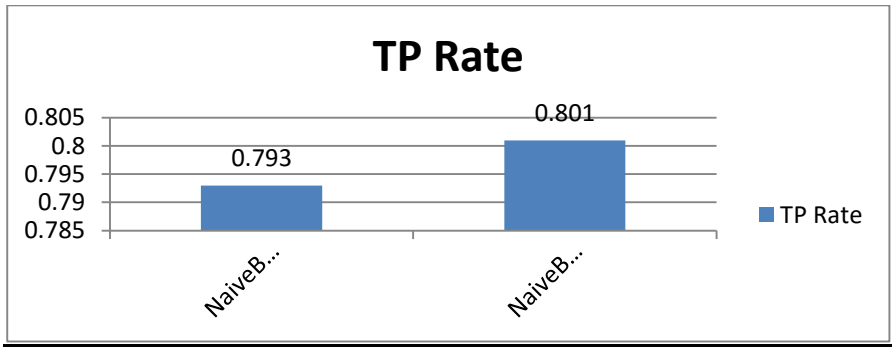

Figure 2: This figure show the TP Rate of existing and proposed classifier The TP Rate of proposed classifier is high as compared to the existing classifier.

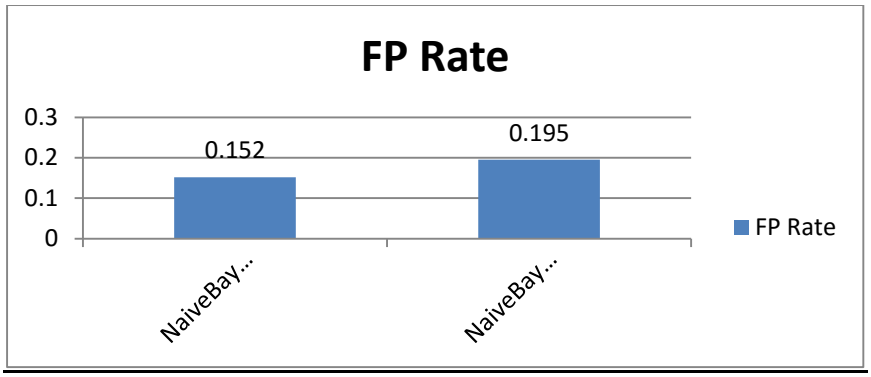

Figure 3: This figure shows the FP Rate of existing and proposed classifier The FP Rate of proposed classifier is high as compared to the existing classifier.

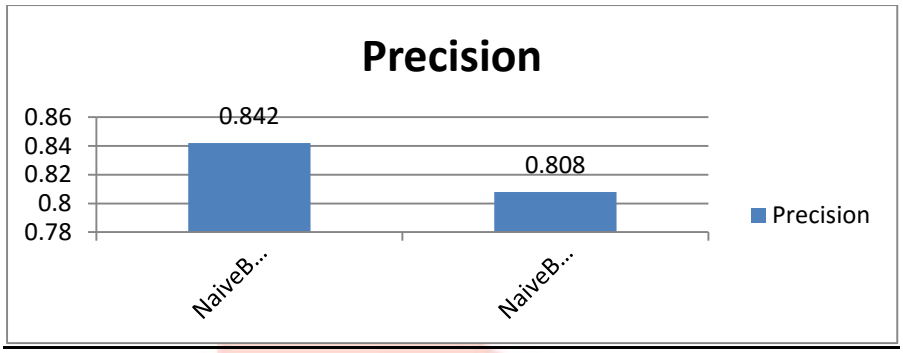

Figure 4: This figure shows the precision of existing and proposed classifier. The precision of existing classifier is low as compared to the proposed classifier.

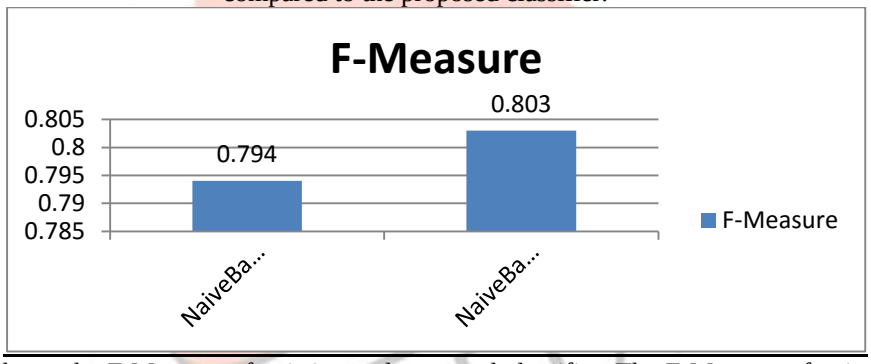

Figure 5: This figure shows the F-Measure of existing and proposed classifier. The F-Measure of existing classifier is low as compared to the proposed classifier.

V. CONCLUSION

We are used the Multi-Naïve Bayes with variable Byte sequence classifier. And multi-Naïve Bayes classifier is used for detection of new malicious executables such that email attachment. This classfier is very popularly and efficiently used data mining method.

The growth in high-speed Internet connections helps malware to spread and infect hosts very rapidly. Consequently it is most important to detect and delete new malware in an effective manner. In this, we propose three data mining algorithms to produce new classifiers with separate features: BayesNet, Naïve – Bayes and a Multi Classifier system and the comparison between three methods. It comprises of root kit data collection, data pre-processing, and classification and performance evaluation phases.

Thus, the proposed approach is implemented and better results have been obtained in order to make the running algorithms better in both time and space.

VI. FUTURE SCOPE

Future aspects for the proposed system are bright. We can use the concept of the malicious detection in email using the different classification algorithm. And we are used multi-Naïve Bayes classifier with variable Byte sequence to detection the new malicious executables like email attachment.

We will implement online spam email detection .and the existing classifier is used the offline spam email detection..

REFERENCES

[1] Suvendu Jena et.al "Security Applications for Malicious Code Detection Using Data Mining", Volume 3 Issue 1, 2015.

[2] Anirudh Harisinghaney et.al, "Text and Image Based Spam Email Classification using KNN, NaIve Bayes and Reverse DBSCAN Algorithm", (2014)
[3] (Somayeh Soltani et.al., "A Survey On Real World Botnets And Detection Mechanisms", 2014)
[4] (Rajkumar E.V.1 Aravindharamanan, "A state of the art review on various malware detection and analysis in web security", 2014)
[5] (Kirti Mathur and Saroj Hiranwal, "A Survey on Techniques in Detection and Analyzing Malware Executables", 2013)
[6] (Anshul Goyal and Rajni Mehta., "Performance Comparison of Naïve Bayes and J48 Classification Algorithms", 2012)
[7] (Ammar Ahmed E. Elhadi, Mohd Aizaini Maarof and Ahmed Hamza Osman, "Malware Detection Based on Hybrid Signature Behaviour Application Programming Interface Call Graph", 2012)
[8] Parisa Bahraminikoo, "Utilization Data Mining to Detect Spyware" IOSR Journal of Computer Engineering (IOSRJCE) Volume 4, Issue 3, pp.01-04, 2012.
[9] (Jianyong Dai et.al., "Efficient Virus Detection Using Dynamic Instruction Sequences", 2009)
[10] Bhavani Thuraisingham et.al, "Data Mining for Security Applications", IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, 2008)