Performance and Comparison of Naive Bayes Classification in WEKA Environment

Richa [*]
Research Scholar Galaxy Global Imperial Tech.Campus
Department of Computer Sc & Engineering
Kurukshetra University
Munjal.richa03@gmail.com

Saurabh Mittal
Associate Professor Galaxy Global Imperial Tech.Campus
Department of Computer Sc & Engineering
Kurukshetra Universty
saurabh. mittal @galaxyglobaledu.com

Abstract-
In our days, electronics attacks can cause a very destructive damage for nations which make necessary the use of completed security policy to minimize the potential threats. Intrusion Detection System (IDS) is becoming a vital research proposed IDS by integrated signature based (Snort) with anomaly based (Naive Bayes) to enhance system security to identify attacks. This research used Knowledge Discovery Data Mining (KDD) CUP 20 Percent dataset and Waikato Environment for Knowledge Analysis (WEKA) program for testing the proposed hybrid IDS. Accuracy, finding rate, time to build model and false alarm rate were used as parameters to evaluate performance with Naïve Bayes, Snort by way of J48graft and Snort with Bayes Net. The result shows good presentation of using Naive Bayes algorithm.

Keywords: NIDS, HIDS, Clustering, Data Set

1. INTRODUCTION

An **intrusion detection system** (**IDS**) is a device or software application that monitors network or system activities for malicious activities or policy violations and produces reports to a management station. IDS come in a variety of “flavors” and approach the goal of detecting suspicious traffic in different ways. There are network based (NIDS) and host based (HIDS) intrusion detection system. Some systems may try to stop an intrusion attempt but this is neither required nor expected of a monitoring system. Intrusion detection and avoidance systems (IDPS) are primarily focused on identifying possible incident, sorting information about them, and reporting attempts. They use several response techniques, which involve the IDPS stop the attack itself, changing the security environment (e.g. reconfiguring a firewall) or changing the attack's content.

For example:- **burglar Alert/Alarm:** A signal suggesting that a system has been or is being attacked.

Data mining can improve variants finding rate, control false alarm rate, and reduce false dismissals. Data mining based on intrusion detection systems can be roughly categorized into major two groups: misuse detection and anomaly detection. Network intrusion detection is the process of monitoring the events occurring in a computing system or network and analyzing them for signs of intrusions, defined as attempts to cooperate the confidentiality. The intrusion attacks can be divided into four categories: Probe (e.g. IP sweep, vulnerability scanning), denial of service (DoS) (e.g. mail bomb, UDP storm), user-to-root (U2R) (e.g. buffer overflow attacks, root kits) and remote-to-local (R2L) (e.g. password guessing, worm attack).

Intrusion detection (ID) is a type of security management system for computers and networks. An ID system gather and analyze in order from various areas within a computer or a network to identify possible security breaches, which include both intrusions (attacks from outside the organization) and misuse (attacks from within the organization). ID uses *vulnerability assessment* (sometimes referred to as *scan*), which is a technology developed to assess the security of a computer system or network.

Intrusion detection functions include:

- Monitoring and analyze both user and scheme activities
- analyze scheme configurations and vulnerabilities
- Assessing system and file integrity
- Ability to identify patterns typical of attacks

- Analysis of anomalous activity patterns
- Tracking user policy violations

Typically, an ID system follows a two-step process. The first events are host-based and are considered the *passive* component, these include: check of the system's configuration files to detect inadvisable settings; inspection of the password files to detect inadvisable passwords; and inspection of other system areas to detect policy violation. The second events are network-based and are considered the *active* component: mechanisms are set in place to reenact known methods of attack and to record system responses.

1.1.1 Network Intrusion detection systems (NIDS) and Host Intrusion detection systems (HIDS)

Network Intrusion Detection Systems (NIDS) usually consists of a network appliance (or sensor) with a Network Interface Card (NIC) operating in promiscuous mode and a separate management edge. The IDS is placed along a network segment or boundary and monitors all traffic on that segment.

A Host Intrusion Detection Systems (HIDS) and software applications (agents) installed on workstations which are to be monitor. The agent monitor the operating system and write data to log files and/or trigger alarms. A host Intrusion detection systems (HIDS) can only monitor the individual workstations on which the agents are installed and it cannot monitor the whole network. Host based IDS systems are used to monitor any intrusion attempts on critical servers.

The drawback of Host Intrusion Detection Systems (HIDS) are

• complex to analyse the intrusion attempts on multiple computers.

• Host Intrusion Detection Systems (HIDS) can be very difficult to maintain in large networks with different operating systems and configurations

• Host Intrusion Detection Systems (HIDS) can be disabled by attackers after the system is compromised.

1.1.2 Knowledge-based (Signature-based) IDS and behavior-based (Anomaly-based) IDS

A knowledge-based (Signature-based) Intrusion Detection Systems (IDS) references a database of previous attack signatures and known system vulnerabilities. The meaning of word name, when we talk about Intrusion Detection Systems (IDS) is recorded evidence of an intrusion or attack. Each intrusion leaves a footprint following (e.g., nature of data packets, unsuccessful attempt to run an application, fail logins, file and folder access etc.). These route are called signatures and can be used to identify and prevent the same attacks in the upcoming. Based on these signatures Knowledge-based (Signature-based) IDS identify intrusion attempts.

The drawback of Signature-based Intrusion Detection Systems (IDS) is signature database must be continually updated and maintained and Signature-based Intrusion Detection Systems (IDS) may fail to identify a unique attacks.

A Behavior-based (Anomaly-based) Intrusion Detection Systems (IDS) references a baseline or learned prototype of normal system activity to identify active intrusion attempts. deviation from this baseline or pattern cause an alarm to be triggered. upper false alarms are often related with Behavior-based Intrusion Detection Systems (IDS).

1.1.3 Clustring Technique

Clustering techniques can be categorized into the following classes: pairwise clustering and innermost clustering. Pairwise clustering (i.e., similarity based clustering) unifies similar data instances based on a data-pairwise distance measure. On the other hand, central cluster, also called centroid-based or model-based clustering, models each cluster by its “centroid”. In terms of runtime intricacy, centroid-based clustering algorithms are more efficient than similarity-based clustering algorithms

II. Related work

Mahbod Tavallaee et al[2009] anomaly detection has attracted the attention of many researchers to overcome the weakness of signature-based IDSs in detecting novel attacks, and KDDCUP’99 is the mostly broadly used data set for the evaluation of these systems. we have proposed a new information set, NSL-KDD, which consists of selected records of the complete KDD dataset and does not suffer from any of mentioned shortcomings. The new version of KDD statistics set, NSL-KDD is publicly available for researchers through our website1. even if, the data set still suffers from some of the problems and may not be a perfect representative of existing real networks, because of the require of public data sets for network-based IDSs, we judge it still can be applied as an effective benchmark data set to help researchers compare different intrusion detection methods.

M.Govindarajan et al[2009] Data Mining is the use of algorithms to extract the information and patterns derived by the knowledge discovery in databases process. Data Mining has become a very useful technique to reduce information overload and improve decision making by extracting and refining useful knowledge through a process of searching for relationships and patterns from the extensive data collected by organization. Data mining technologies, such as rule training, neural networks, genetic algorithms, fuzzy logic, and irregular sets are used for classification and pattern recognition in many industries.

Anomaly detection assumes that an intrusion will always reflect some deviations from normal patterns. Anomaly detection may be divided into static and dynamic detection. A static anomaly detector is based on the assumption that there is a portion of the system being monitored that does not modify. Usually, static detectors only address the software portion of a system and are based on the assumption that the hardware need not be checked. The stationary portion of a system is the code for the system and the constant portion of data upon which the correct functioning of the classification depends. For example, operating systems software and data to bootstrap a computer never change. Dynamic detection typically operates on audit records or on monitored networked traffic data. Audit records of operating systems do not record all events that is recorded in the audit will be observed and these events may occur in a sequence. Misuse detection is based on the knowledge of system vulnerabilities and known attack pattern. Misuse detection is disturbed with finding intruders who are attempting to break into a system by exploiting some known vulnerability.

Tayeb Kenaza et al[2010] we propose to combine a behavioral intrusion detection approach with a clustering approach in order to obtain a set of clusters with different false alerts rates. The classify of these clusters with respect to their false alerts rates will be considered as an alerts prioritization. therefore, new alerts will be classified to the closest cluster and processed according to their cluster priority. Several machine learning mechanisms such as neural networks, support vector machines (SVM), decision trees, Bayesian networks, etc. have been used for the design of behavioral based intrusion detection systems (IDS).

Chang-Tien Lu,et al[2005] Security is becoming a critical part of organizational information systems. Intrusion Detection System (IDS) is an important detection that is used as a countermeasure to preserve data integrity and system availability from attack. Data mining is being use to fresh, classify, and study large amount of network data to correlate common infringement for intrusion detection. Several Data Mining techniques such as clustering, classification, and involvement rules are proving to be useful for gathering different knowledge for Intrusion Detection. Computer Security is the ability to protect a computer system and its resources with respect to confidentiality, integrity, and availability. Various protocols, firewalls are in existence to protect these systems from computer threats. Intrusion is a type of cyber attack that attempts to bypass the security mechanism of a computer system. Such an defender can be an outsider who attempts to access the system, or an insider who attempts to gain and misuse non-authorized privileges.

Intrusion Detection System (IDS) is an important detection used as a countermeasure to preserve data integrity and system availability from attacks. Intrusion Detection Systems (IDS) is a combination of software and hardware that attempts to perform intrusion detection. Intrusion detection is a process of congregation intrusion related knowledge occurring in the process of monitoring the events and analyzing them for sign or intrusion. It raise the alarm when a probable intrusion occurs in the system.

Meng Jianliang et a[2009l Internet security has been one of the most important problems in the globe. Anomaly detection is the basic method to defend new attack in Intrusion Detection .Network intrusion detection is the process of monitoring the events occurring in a computing system or network and analyzing them for signs of intrusions, defined as attempts to compromise the privacy. Clustering is the method of grouping objects into meaningful subclasses so that the members from the same cluster are moderately similar, and the members from different clusters are quite different from each other. Until now, the clustering algorithms can be categorized into four main groups: partition algorithm, hierarchical algorithm, density-based algorithm and grid-based algorithm. Partitioning algorithms construct a partition of a database of N objects into a set of K clusters.

Z. Muda et al[2011] Intrusion Detection System (IDS) plays an effective way to achieve higher security in detecting malicious activities for a pair of years. Anomaly detection is one of intrusion detection system. Current anomaly detection is frequently associated with high false alarm with moderate accuracy and detection rates when it's unable to detect all types of attacks correctly. To overcome this problem, we recommend an hybrid learning approach through combination of K-Means clustering and Naïve Bayes arrangement. With the rapid growth of network technology, a cyber crime incident has also grown accordingly. A wide range of risks and threats against uncontrolled and undefended assets such as database and web server as well as entire network system become the general concern for intruders nowadays. Gaining unauthorized access to files, network and any other serious security threat can be detected by employing Intrusion Detection System. IDS identify any activity that violates the security policy from various areas within computer and network environment.

There are two traditional IDSs used to detect intruders: signature-based detection and anomaly-based detection. A signature-based IDS match define signature with each anayzed packets on the network to detect known malicious attack as a same way like a virus scanner. These type of IDS required a frequent updating for the new signatures to keep the signature database up-to date Thus, it fails in discovering and detect an unknown attacks once the signature did not exist in its library. Unlike signature-based detection, anomaly-based detection is designed to capture any activities which are diviates the normal usage pattern called normal profile. In recent years, data mining approach have been proposed and used as detection techniques for discover unknown attacks. This approach has resulted in high accuracy and good detection rates but with moderate false alarm on novel attacks.

Slobodan Petrovic et al[2006] Intrusion detection systems (IDS) are security tools designed to detect and classify attacks against computer networks and hosts. They can activate in two ways: either by searching for specific patterns in data (misuse based IDS) or by recognising certain deviations from expected behavior (anomaly based IDS). In anomaly based IDS, clustering algorithms are often used for recognition of ”abnormal” behaviour. They can be applied either directly on incoming data or as a supporting technique in a stage posterior to data classification performed by means of other techniques. Anomaly based IDS classify input data into a number

of categories, or classes. This number can be random, but as the necessary goal of these systems is to distinguish between ”normal” and ”abnormal” behaviour, it is very common to partition the incoming resource access requests into two classes that correspond to these two types of behavior.

Cuixiao Zhang et al[2009] Through analyzing the advantages and disadvantages between anomaly detection and misuse detection, a mixed intrusion detection system (IDS) representation is designed. First, data is examined by the misuse finding module, then abnormal data detection is examined by anomaly detection module. With rapid development of Internet, computer and computer network becomes the attack target of computer crime gradually. It causes company image damaged, the business information stolen, the data robbed and the emollient competition opportunity lost etc, information security becomes one of the most important research issues in the field of information technology. Traditional operate system's reinforcing technique and fire wall technique are all static security defense technologies, which lack active reaction to attack. As a kind of active security technique, intrusion detection is a rational supplement to the traditional static defense technologies

Wireless Communication System is playing a big important role in information systems and its management is very important and vital. There are many managements such as composition management, fault management, presentation management, security management, accounting management and etc. Fault management is more important.

A node becomes faulty because of battery discharge, crash and limitation in age. The presence of faulty node affects the efficiency and throughput of the network, which makes the network inconsistent. Faultynodes cannot communicate with the other mobiles or behave unexpectedly andsend unexpected results. Thus it unnecessarly consumes energy and causeinconsistency.

Abhay Kumar et al[2012] Two common data mining techniques for finding hidden patterns in data are clustering and arrangement analysis. Classification is supposed to be supervised learning and clustering is an unsupervised classification with no predefined module. Clustering tries to group a set of objects and find whether there is some relationship between those items. In this paper we have used the numerical results generate throughout the Probability Density Function algorithm as the basis of recommendations in favor of the K-means clustering for weather-related predictions.

Mouaad KEZIH et al [2013] Intrusions detections systems from point of view of security policy are a second line of defense; they have a supervisory role to observe the activities of our network or hosts to identify attacks in actual time. In our days, electronics attacks can cause a very destructive damage for nations which make necessary the use of completed security policy to minimize the probable threats. IDS it is a very important element to resist against this vulnerability, (KDD) CUP 99 and a Data Mining Tools Waikato Environment for Knowledge Analysis (WEKA) to combine the advantages of an intrusion detection algorithm (PART) and two techniques of Dimensionality Reduction(best first search and genetic search), to estimate our works, we applied the proposed combined technique ,and we check the results by using a several evaluations parameters.

iii.Data Set

Data required for training the unsupervised anomaly based intrusion detection system is taken from the ‘kddTrain 20 percent dataset’. It is assumed that the relative amount of attacks in the training data is very small compared with usual data, a reasonable assumption that may or may not hold in the real world context for which it is useful. If this assumption holds, anomalies and attacks may be detected based on cluster size. big clusters correspond to usual data, and small clusters.

The training data consists of approximately 50,000 normal connection records extracted randomly from the kddcup data set and all the attack connection records with a maximum of 1% per attack type. Thus a training data file with around 50,000 normal instances and 5000 attack instance is obtain. The hard data file is created in the same manner from the correct file. It contain approximately 20,000 normal connection records and around 5000 intrusive connection report. Thus, we are training the anomaly detection system on some number of attacks and testing it to check whether it is able to detect DOS unseen attacks which exist in the testing file. The performance of intrusion detection systems of detection rate, accuracy, false alarm and time build to model. Accuracy can be defined as a percentage of the connections that classified correctly over the whole connections. The proportion of detected attacks data called Detection rate, while indicates the amount of normal data which is falsely detected an attack.

Weka data mining tools [9] were used to generate naïve Bayesian and J48 classifiers with default settings and five-fold cross-validation. The outcome are shown in the following tables. Table 3.1 shows the performance of NaIve Bayesian (NB) classifier, and Table 3.2 shows its confusion matrix. It is wellrecognized in data mining that the following measures provide more informative evaluation of classifier performance when dealing with class-imbalanced data: recall, precision (prec.), F-measures, sensitivity, and specificity [5], which are defined as]

TP is the number of positive cases classified correctly, FN is the number of positive cases classified as negative, FP is the number of negative cases classified as positive, and TN is the number of negative cases classified correctly. In Table3.2 , the TP is 10298, FN is 1445 FP is 1177 and TN is 12272

	TP Rate	**FP Rate**	**Prec**	**Recall**	**F-Measure**	**Roc Curve**	**Class**
	0.912	0.123	0.895	0.912	0.903	0.968	Normal
	0.877	0.088	0.897	0.877	0.887	0.963	Anomaly
WAvg	0.896	0.106	0.896	0.896	0.896	0.966	

Table3.1: NB Output From Weka

a	**b**	**Classified as**
12272	1177	Normal
1445	10298	Anomaly

Table3.2Confusion Matrix NB

	TP Rate	**FP Rate**	**Prec**	**Recall**	**F-Measure**	**Roc Curve**	**Class**
	0.991	0.064	0.947	0.991	0.969	0.996	Normal
	0.936	0.009	0.989	0.936	0.962	0.996	Anomaly
WAvg	0.996	0.038	0.967	0.966	0.966	0.996	

Table3.3: Bays Net Output From Weka

a	**b**	**Classified as**
13330	119	Normal
747	10996	Anomaly

Table 3.4:Confusion Matrix Bays Net

	TP Rate	**FP Rate**	**Prec**	**Recall**	**F-Measure**	**Roc Curve**	**Class**
	0.998	0.002	0.998	0.998	0.998	1	Normal
	0.998	0.002	0.998	0.998	0.998	1	Anomaly

W Avg	0.998	0.002	0.998	0.998	0.998	1	

Table3.5: J48Graft Output From Weka

a	b	Classified as
13425	24	Normal
25	11718	Anomaly

Table 3.6Confusion Matrix J48 Graft

Accuracy = (TN+TP/ (TN+TP+FN+FP))

recall = TP/(TP+FN)

F-measure = (2*recall *precision) I (recall + precision),

sensitivity = TP/(TP+FN) = recall,

specificity = TN/(FP+ TN).

	Native Bays	Bays Net	J48Graft
Accuracy	89.59	96.5	99.8
Dtection Rate	87.69	93.6	97.8
False Alarm Rate	8.75	0.88	0.17
Time to Build Model	5.75	6.53	28.53

FiG 3.7Comparison of NB, Bays Net, and J48 Graft

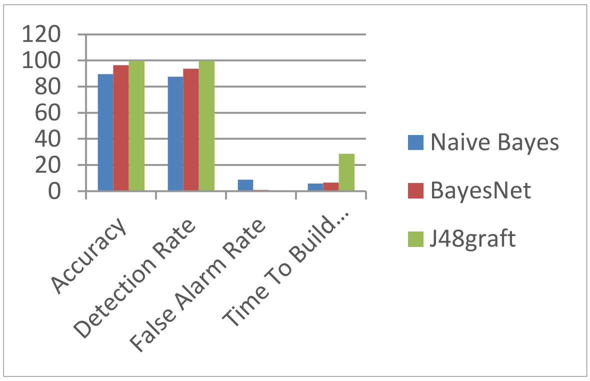

Fig3.8:Comparision between three algorithm

1Accuracy:

Accuracy is the proportion of correct class namely True Positive (TP) and True Negative total number of classifications [5]. This research accuracy rate based on formula (1) which that diagonal elements in the confusion matrix elements in the confusion matrix.

Accuracy = (TN+TP/ (TN+TP+FN+FP)) (1)

2. Detection Rate

The Total Detected attack amongst all the scanned data is called detection rate

Detection rate = TP/ (TP+FN) (2)

3. False alarm rate

False alarm rate is the proportion of normal data which is falsely detected and labeled as an attack, namely False Positive (FP) over the sum of False Positive and True Negative(TN) elements multiply by hundred[5]. False alarm rate calculated based on formula (3)

False alarm rate= FP/ (FP+TN)*100 % (3)

4. Time to build model

Time taken for each algorithm to build a model had been observed and recorded. The time unit used in this research is in minute.

CONCLUSIONS

Data mining in intrusion detection is a relatively new perception. Thus there will likely be obstacles in developing an effective solution Intrusion detection systems have been an area of active research for over fifteen years. Current commercial intrusion detection systems employ misuse detection. As such, they entirely lack the ability to detect new attacks. it is impossible to prevent security violation completely by using the exciting security technologies. Accordingly, Intrusion Detection is an important component of network security.

Compare the rate of accurate and inaccurate classified instance between all three algorithm .it can be seen clearly that rate of accuracy, detect rate and false alarm rate with run J48 Graft show better result than the other two algorithm.

REFERENCES

1.Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani "**A Detailed Analysis of the KDD CUP 99 Data Set"** Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009).

2M.Govindarajan,Rlvl.Chandrasekaran "**Intrusion Detection Using k-Nearest Neighbor"** Volume 2 Issue 3 IEEE 2009

3.Tayeb Kenaza, Abdelhalim Zaidi "**Clustering approach for false alerts reducing in behavioral based intrusion detection systems"** Volume 4 Issue 5 IEEE 2010.

4. Chang-Tien Lu, Arnold P. Boedihardjo, Prajwal Manalwar "**Exploiting Efficient Data Mining Techniques to Enhance Intrusion Detection Systems"** Volume 2 Issue 3 IEEE 2005

5. Meng Jianliang Shang Haikun Bian Ling "**The Application on Intrusion Detection Based on K-means Cluster Algorithm"** IEEE 2009 International Forum on Information Technology and Applications.

6. Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir "**Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification"** 2011 7th International Conference on IT in Asia (CITA).

7. Slobodan Petrovic, Gonzalo A´ lvarez2, Agust´ın Orfila3, and Javier Carbo "**Labelling Clusters in an Intrusion Detection System Using a Combination of Clustering Evaluation Techniques"** Proceedings of the 39th Hawaii International Conference on System Sciences – 2006.

8. Cuixiao Zhang; Guobing Zhang; Shanshan Sun "**A Mixed Unsupervised Clustering-based Intrusion Detection Model"** 2009 Third International Conference on Genetic and Evolutionary Computing.
9. Abhay Kumar, Ramnish Sinha **"Modeling using K-Means Clustering Algorithm"** 1st Int'l Conf. on Recent Advances in Information Technology | RAIT-2012IEEE .
10. Mouaad KEZIH Mahmoud TAIBI**" Evaluation Effectiveness of Intrusion Detection System with Reduced Dimension Using Data Mining Classification Tools**"2013 2nd International Conference on Systems and Computer Science (ICSCS) Villeneuve d'Ascq, France, August 26-27, 2013