IDS Using Classification Teachniques in Weka Environment With Feature Reduction

Jyoti Deshwal[1] Savita Gupta[2]
[1]Assisatnt Professor
[1]Department of Computer Science Engineering
[1]SKIET Kurukshetra University, Kurukshetra,Haryana, India

Abstract— In this paper, Intrusion Detection System (IDS) based on combining data mining is presented and implemented in WEKA. this exposure, in our works, we use a wired data base Knowledge Discovery Data Mining (KDD) CUP 10Percent and a Data Mining Tools Waikato Environment for Knowledge Analysis (WEKA) we check the results by using a several evaluations parameters. The results illustrate that a very high detection rate for certain attacks types. we proposed to calculate the mean value via sampling different ratios of normal data for each measurement, which lead us to reach a better accuracy rate for observation data in real world This presents useful information in intrusion detection.

Key words: IDS, Data Mining, Attack, Clustrin, Weka

I. INTRODUCTION

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.

It is a process of gathering intrusion related knowledge occurring in the process of monitoring the events and analyzing them for sign or intrusion. It raises the alarm when a possible intrusion occurs in the system. The network data source of intrusion detection consists of large amount of textual information, which is difficult to comprehend and analyze. The main motivation behind using intrusion detection in data mining is automation. Pattern of the normal behavior and pattern of the intrusion can be computed using data mining.

To apply data mining techniques in intrusion detection, first, the collected monitoring data needs to be preprocessed and converted to the format suitable for mining processing.

Next, the reformatted data will be used to develop a clustering or classification model. The classification model can be rule-based, decision-tree based, association-rule based, Bayesian-network based, or neural network based. Intrusion Detection mechanism based on IDS are not only automated but also provides for a significantly elevated accuracy and efficiency. Unlike manual techniques, Data Mining ensures that no intrusion will be missed while checking real time records on the network. Credibility is important in every system.

IDS are now becoming important part of our security system, and its credibility also adds value to the whole system. Data mining techniques can be applied to gain insightful knowledge of intrusion prevention mechanisms. They can help detect new vulnerabilities and intrusions, discover previous unknown patterns of attacker behaviors, and provide decision support for intrusion management. Intrusion detection is detection of intrusion behavior, it collects information of the key part of computer network and system, then analyzes them to detect whether occur the

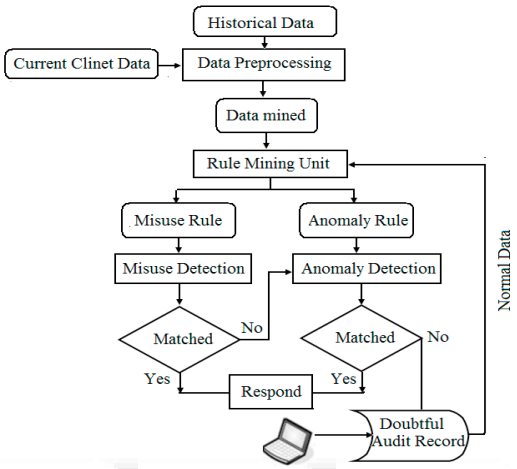

Fig. 1.1: Data Mining System Structure in IDS

action of disobey security strategy. Intrusion Detection System (IDS) is the software or combination of software and hardware to detect intrusion behavior. IDS can examine intrusion attack before system is damaged, and make use of alerting and defense system to deport the intrusion attack. In the process of intrusion attack, It can reduce the loss resulted in. After system attacked, the related attack information is collected, and as security system knowledge, it is added to the strategy set, thus can strengthen system security defence ability, avoid system being intruded by the same intrusion again.

II. TYPES OF ATTACKS

The simulated attacks fall in one of the following four categories:

A. Denial of Service Attack (DoS)

is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine.

B. User to Root Attack (U2R)

is a class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.

C. Remote to Local Attack (R2L)

occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.

D. Probing Attack

is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

III. KDD CUP 99 DATA SET DESCRIPTION

Since 1999, KDD'99 has been wildly used data set for the estimate of anomaly detection methods is built based on the data captured in DARPA'98 IDS evaluation program. DARPA'98 is about 4 gigabytes of compressed raw (binary) tcp dump data of 7 weeks of network traffic, which can be processed into about 5 million connection records, each with about 100 bytes. The two weeks of test data have around 2 million connection records. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack.

The replicated attacks fall in one of the following four categories: User to Root ; Remote to Local; Denial of Service; and probe

The datasets enclose a total number of 24 training attack type, with an extra 14 types in the test data only. KDD'99 features can be classified into three groups:

A. Basic features

this category encapsulates all the attribute that can be extracted from a TCP/IP connection. Most of these features are important to an implicit delay in detection.

B. Traffic features

this group includes features that are computed with respect to a window interval and is divided into two groups:

(1) "Same host" features: inspect only the connections in the past 2 seconds that have the same target host as the existing relation, and calculate statistics related to protocol behavior, service, etc.

(2) "same service" features: inspect only the connections in the past 2 seconds that have the same service as the existing connection

C. Content features

unlike most of the DoS and Probing attacks, the R2L and U2R attacks don't have any intrusion frequent sequential patterns. This is because the DoS and Probing attacks involve many connections to some host(s) in a very short period of time; however the R2L and U2R attacks are embedded in the data portions of the packets, and normally involves only a single connection. To detect these kinds of attacks, we need some features to be able to look for suspicious behavior in the data portion, e.g., number of failed login attempts. These features are called content features.

IV. DATA MINING TECHNOLOGY

Data mining is the latest introduced technology of intrusion detection. Its advantage lies in the fact that it can withdraw the needed and unknown knowledge and regularities from the massive network data and host log data. It is a new attempt to use data mining in achieving network security, both at home and abroad. At present, data mining algorithm applied to intrusion detection mainly has four basic patterns: association, sequence, classification and clustering. Data mining technology is advanced for:

(1) It can process large amount of data.

(2) It doesn't need the users' subjective evaluation, and is more likely to discover the ignored and hidden information.

These two are especially applicable to the intrusion detection based on analyzing the abnormality of auditing record.

A. WEKA

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied
directly to a dataset or called from your own Java code . Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. WEKA consists of Explorer, Experimenter, Knowledge flow, Simple Command Line Interface, Java interface,

The WEKA tool incorporates the following steps :

- Analysis and pre-processing of the features in the database and assessing the correctness of the data.
- Definition of the class attributes which divide the set of instances into the appropriate classes.
- Extraction of the potential features to be used for classification.
- Selection of a subset of features to be used in the learning process.
- Investigation of a possible imbalance in the selected data set and how it may be counteracted.
- Selection of a subset of the instances, i.e. the records that learning is to be based on.
- Application of a classifier algorithm for the learning process.
- Decision on a testing method to estimate the performance of the selected algorithm.

V. PURPOSED METHOD

The performance of intrusion detection systems depends upon the basic factors like 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 were used to generate naïve Bayesian and J48 classifiers with default settings and five-fold cross-validation

A. Feature Selection Process

There are many factors that affect the success of data mining algorithm on a given task. The quality of the data is one such factor- if information is irrelevant or redundant , or the data is noisy or unreliable, then knowledge discovery during training in more difficult. There are two methods through which we can perform feature selection process:

1) Wrapper Method

This will create all possible subset from our feature vector. Then it will use a classification algorithm to induce classifier from the features in each subset. It will consider the subset of features with which the classification algorithm performs the best. To find a subset the evaluator will use a

search technique(random search, BFS, DFS & hybrid search).

2) Filter Method

This use an attribute evaluator and a ranker to rank all the features in a data set. The number of features we want to select from feature vector can always be defined. In this method we need to omit the data one at a time that have lower ranks and check the predictive accuracy of the classification algorithm.

B. Naïve Bayes Classifier

The naïve Bayes model is a heavily simplified Bayesian probability model.

In this model, consider the probability of an end result given several related evidence variables. The probability of end result is encoded in the model along with the probability of the evidence variables occurring given that the end result occurs. The probability of an evidence variable given that the end result occurs is assumed to be independent of the probability of other evidence variables given that end results occur. analyzes the relationship between independent variable and the dependent variable to derive a conditional probability for each relationship.

Using Bayes Theorem we write:

$$P(H|X) = P(X|H)\,P(H) / P(X)$$

Let X be the data record. Let H be some hypothesis represent data record X, which belongs to a specified class C. For classification, we would like to determine P(H|X), which is the probability that the hypothesis H holds, given an observed data record X. P(H|X) is the posterior probability of H conditioned on X. In contrast, P(H) is the prior probability. The posterior probability P(H|X), is based on more information such as background knowledge than the prior probability P(H), which is independent of X. Similarly, P(X|H) is posterior probability of X conditioned on H.

Bayes theorem is useful because it provides ways to calculate the posterior probability P(H|X) from P(H), P(X),and P(X|H)

C. J48 Algorithm

J48 are one type of decision tree. It is an optimized version of C4.5 algorithm. When any specific data item is classified, it will be divided in different levels starting from root node to the leaf or terminal node in a hierarchical manner. This process will continue until it reaches over the terminal node which further cannot be subdivided. This Decision Tree are used in decision analysis, in this tree every non-leaf node represents a test or decision on the data item. Depending upon output at each level it will choose certain branch. For example a question has multiple answers, and each answer can further be divided, it will subdivide up to the last level. Decision Tree is a very powerful technique which is used for real world problems by classified the problem into tree structure and applies the control rules over the internal nodes

D. Bayes Net

By using the bayes theorem BayesNet can be developed. To structure a Baysian network first conditional probability of every node must be calculated. Acyclic graphs are used to represent the network. Before building the network, it is assumed that there are no missing values and all attribute values are nominal. Different types of estimators (BayesNet Estimator, BMA Estimator, Multi Nominal BMA Estimator, Simple Estimators) and algorithms (Genetic Search, Hill climber, K2, LAGD Hill climber, Repeated Hill climber, Simulated Annealing, Tabu Search, TAN) were used to estimate the probability. The output was visualized by using graph.

Now, Weka data mining tools is used to generate naïve Bayesian and J48 classifiers with default settings and five-fold cross-validation. The results are shown in the following tables. Table 1 shows the performance of NaIve Bayesian (NB) classifier, and Table 2 shows its confusion matrix. It is well recognized 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, which are defined as:

- **TP**-True Positive indicate the number of attack record that are correctly classified.
- **TN**- True Negative indicate the number of valid record that are correctly classified.
- **FP**-False Positive indicates record that are incorrectly classified as attack.
- **FN**- False Negative indicates record that are incorrectly classified as valid activities where in fact they are attack.

E. Infogain Attribute Eval with Ranker

The original dataset consist of 41 attributes and one class label. InfoGainAttribute Eval in Weka decide which attribute is the best using a statistical property, called information gain, is used. Gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information (information being the most useful for classification) is selected with ranker that provides the rank number to all the attributes on the basis of their priority. The total number of attributes that are left after applying feature selection are 36 and the given below are the confusion matrix obtained after applying J48graft, Bayes Net and Naïve Bayes Classifier on these attributes

A	b	Classified as
12269	1180	Normal
1416	10327	Anomaly

Table 5.1: Confusion Matrix NB

A	B	Classified as
13325	124	Normal
747	10996	Anomaly

Table 5.2: Confusion Matrix Bays Net

A	B	Classified as
13391	58	Normal
66	11667	Anomaly

Table 5.3: Confusion 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

Comparison of NB, Bays Net, and J48 Graft

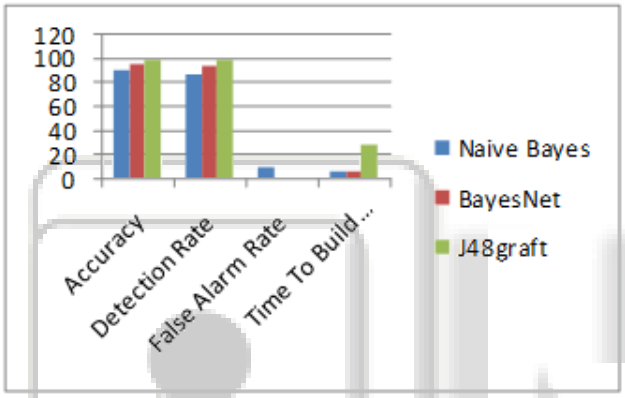

Fig. 5.1: Comparision between three algorithm

1) Accuracy:

Accuracy is the proportion of correct class namely True Positive (TP) and True Negative total number of classifications [3]. 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[3]. 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.

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

VI. Conclusions

The researcher found that there are differences between uses each of the algorithms Naïve Bayes, J48graft and Bayes Net, in term of accuracy, detection rate, false alarm rate and taken time to build model. In some strategies such as Percent Correct Classification (PCC) accuracy, the percentage of correctly classified instances in J48graft was higher than percentage of correctly classified instances with uses bayes net and naïve bayes respectively. While, the time taken to build model by naïve bayes was much faster than taken time to build model by Bayes Net and J48graft respectively. In term of detection rates naïve bayes showed better result comp ares to bayes net and J48graft respectively. Finally J48graft performed better result compares to Bayes Net and Naïve Bayes respectively in term of false alarm.

In sampling, the research supposes that the distribution of attack data other than normal data is even, which cannot surely get optimal results, and this should be improved and validated in future.

References

[1] [Mrutyunjaya 07] Mrutyunjaya Panda and Manas Ranjan Patra "Network Intrusion Detection Using Naïve Bayes" Vol. 7, No.3 IJCSNS, December 2007

[2] [Mahbod 09] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani "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).

[3] [Govindrajan 09] M.Govindarajan# 1, Rlvl.Chandrasekaran" Lecturer (Senior Scale), "Intrusion Detection Using k-Nearest Neighbor" Volume 2 Issue 3 IEEE 2009.

[4] [Robu 10] R. Robu and V. Stoicu-Tivadar "Arff Convertor Tool for WEKA Data Mining Software" Proceedings of the IEEE 2010.

[5] [Sanoop 11] Sanoop Mallissery 1, Jeevan Prabhu 2, Raghavendra Ganiga "Survey on Intrusion Detection Methods" Proc. of1nt. Con/, on Advances in Recent Technologies in Communication and Computing 2011.

[6] [Muda 11] 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] [Yang 12] Yang Yong "The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm" Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Hainan University.

[8] [Wankhade 13] Kapil Wankhade, G. H. Raisoni "An Efficient Approach for Intrusion Detection Using Data Mining Methods" Proceedings of the IEEE 2013.

[9] [Chandarasekhar 13] A. M Chandrasekhra; K. Raghuveer, "Intrusion Detection Technique by using K-mean, Fuzzy Neural Network and SVM Classifiers" International conference 2013, IEEE

