ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering
Vol. 2, Issue 7, July 2013

Intrusion Detection in Data Mining With Classification Algorithm

Patel Hemant[1], Bharat Sarkhedi[2], Hiren Vaghamshi[3]

Associate Professor, Dept. of CSE/IT, Dr.Subhash Technical Campus, Junagadh, Gujarat, India[1]

Student, Dept. of CE, Patel College of Science & Technology, Bhopal, Madhya Pradesh, India[2]

Student, Dept. of CE, Nobel College of Engineering, Junagadh, Gujarat, India[3]

ABSTRACT: In the research of intrusion detection there are so manyattacks in the real life and some IDS System to detect it like network-based IDS, host-based IDS and application-based IDS to detect the Intrusion. In this paper, one frame work is introduce to detect an intrusion data with the help of data mining, a framework for intrusion detection system for filtering data set to network attacks. We also talk about the basic data mining technology for finding intrusion data for the data set. Detection in the field of data mining for intrusion detection. We also discuss some of the common algorithms for intrusion detection, such as decision trees, Naive Bayes, Naive Bayes (CFSGSW), NBTree improved adaptive NBTree it.

Keywords: Data mining, Intrusion Detection, Network, Decision Tree, Naive Bayes, NBTree**.**

I.INTRODUCTION

Fast growing of the Internet, network security has become the fundamentals issues of computer technology. So the main task of the technology expert is to provide a secure data confidentiality, data integrity and data availability. Invasion is an action, trying to destroy data confidentiality, data integrity and availability of network information. There are many intrusion detection attacks that have to handle by the security framework or the system security tools. . In fact, intrusion detection is a knowledge which is collected during the invasion of the monitoring and analysis of data to conclude whether the system is not as invasive or user activities occurring in the system, system logs, etc. The possible intrusion detection, IDF sounds the alarm to the network administrator.

Here we have introduce the framework for intrusion detection in data mining which is used to detect an insecure network attacks on computer systems .the detail introduction about the framework is discuss in the next chapter in this paper.

Typically, Intrusion detection system can be classified into three systems based on such (i) misuse based system, (ii) anomaly based systems, and (iii) hybrid systems. Misuse based IDS simple pattern matching techniques to match the attack pattern, and a database of known attack patterns are consistent, and produce very low false positive (FP). It requires the signature of the rules or to see, not so well-known attacks regularly updated. Anomaly based of the IDS to determine the normal behaviour by examining the abnormal behaviour of the new attack [2], both well-known and achievea high detection rate (DR) unknown attacks, but makes many false positives (FP). Anomaly based IDS, the development of IDS audit data collected by observing the rules. Developed by the operating system audit data record of the activities is logged to a file in chronological order. On the other hand, a combination of a hybrid IDS based on misuse and corruption of the detection system technology. The current adaptive intrusion detection is designed to address large amounts of data in the analysis of audit, inspection rules for performance optimization.

On the other hand, according to the resources to track them, the IDS system is classified into three categories: (i) network-based IDS, (ii) host-based IDS and (iii) application-based IDS. Network-based intrusion detection systems to monitor network traffic and use the contents of the original package network to analyse network protocols, transport and application to identify suspicious activity [3]. Host-based IDS monitors a machine and audit tracking data from the host operating system. A common example is the data audit system calls, events, resource utilization, and Windows NT and UNIX environments syslog Note [4]. Application-based IDS can be divided into host-based IDS. It analyses the events in the sweat software applications.

Most IDS detects malicious activity, whether this is at the transaction level or OS or both of the transaction and the standard OS [5-6]. The ID has some disadvantages:

- A data overload. In fact, there are many attacks on the emergence of the privilege. To test these attacks, it is necessary to collect large Amount of information, including system logs, Transaction records, etc., to analyse. However, it is Difficult to understand and traditional IDS Analysis of data collected, due to extensive Data source intrusion detection.
- False positives. Another disadvantage of the traditional IDS is a false positive occurs when the amount of Ordinary attack wrongly classified.
- False negative. This is the IDS system does not alarm when there is a real Attack the system.

The rest of this paper is organized as follows:
Section I we describe the types of attack in dataset. Section II presents the detailed insight of the various data mining algorithms and advantage and disadvantage of the algorithm, in section III give the detail about the proposed framework. While in SectionIV we talk about the future work and conclusions.

II.TYPE OF ATTACK

The classes in KDD99 dataset can be categorized into five main classes (one normal class and four main intrusion classes: probe, DOS, U2R, and R2L) [7-8].

1) Normal connections are generated by simulated daily user behaviour such as downloading files, visiting web pages.
2) Denial of Service (DoS) attack causes the computing power or memory of a victim machine too busy or too full to handle legitimate requests. DoS attacks are classified based on the services that an attacker renders unavailable to legitimate users like apache2, land, mail bomb, back, etc.
3) Remote to User (R2L) is an attack that a remote user gains access of a local user/account by sending packets to a machine over a network communication, which include send-mail, and Xlock.
4) User to Root (U2R) is an attack that an intruder begins with the access of a normal useraccount and then becomes a root-user by exploiting various vulnerabilities of the system. Most common exploits of U2R attacks are regular buffer-overflows, load-module, Fd-format, and fb-config.
5) Probing (Probe) is an attack that scans a network to gather information or find knownVulnerabilities. An intruder with a map of machines and services that are available on a network can use the information to look for exploits.

III.CLASSIFICATION ALGORITHMS

System construction of the classification is commonly used data mining tools. Such a system as input is collected, each belonging to a fixed property settings, and output a classification described in a small class numbers, can accurately predict a new case belongs to [9] class. In this paper, based on Decision Tree, Naive Bayes, Naive Bayes (CFSGSW), NBTree and Improved Self Adaptive NBTree.

A. Decision Tree

Decision trees are well known machine learning techniques. A decision tree is composed of three basic elements:

- An edge or a branch corresponding to the one of the possible attribute values this means one of the test attribute outcomes.
- A decision node specifying a test attributes.
- A leaf which is also named an answer node contains the class to which the object belongs.

In decision trees, two major phases should be ensured: Building the tree. Based on a given training set, is to build a decision tree. It consists of each decision node, select the "appropriate" test properties, and define the class label of each leaf.

Classification Order to classify a new instance; we began to determine the root of the tree, then we test the node specified property. The test results, allowing moving down the tree relative to a given instance of the attribute value.This process is repeated until it encounters a leaf [10]. The instance is then classified in the same class characteristics to leaves.

The attribute selection measure taking into account the discriminative power of each attribute over classes in order to choose the ‘best’ one as the root of the (sub) decision tree. In other words, this measure should consider the ability of each attribute A_k to determine training objects' classes. In the literature many attribute selection measures are proposed[11-12]. We mention the gain ratio, used within the C4.5algorithm [12] and based on the Shannon entropy, where for an attribute A_k and a set of objects T, it is defined as follows:

$$Gain(T, A_k) = Info(T) - Info_{A_k}(T) where$$

$$Info(T) = -\sum_{i=1}^{n} \frac{freq(c_i,T)}{|T|} \log 2 \frac{freq(c_i,T)}{|T|}$$

$$Info_{A_k}(T) = \sum_{a_k \in D(A_k)} \frac{|T_{a_k}^{A_k}|}{|T|} Info(T_{a_k}^{A_k})$$

And freq(ci, T) denotes the number of objects in the setT belonging to the class ci and TAKak is the subset of objectsfor which the attribute Ak has the value ak (belonging tothe domain of Ak denoted D(Ak)).Then, Split Info (Ak) is defined as the information contentof the attribute Ak itself [12]

$$Split\ Info(T, A_k) = -\sum_{a_k \in D(A_k)} \frac{|T_{a_k}^{A_k}|}{|T|} \log 2 \frac{|T_{a_k}^{A_k}|}{|T|}$$

So, the gain ratio is the information gain calibrated by Split Info:

$$Gain\ ratio(T, A_k) = \frac{Gain(T, A_k)}{Split\ Info(A_k)}$$

The partitioning strategy having as objective to divide the current training set by taking into account the selected test attribute.The stopping criteria dealing with the condition(s) of stopping the growth of a part of the decision tree (or even all the decision tree). In other words, they determine whetheror not a training subset will be further divided.

B.Naive Bayes

Naive Bayes (NB) is a method of supervised classification learning commonly used to predict the likelihood of group members. It assumes conditional independence of class, is based on Bayes theorem. Bayesian network is one of the most widely used graphics Model representation and processing of uncertain information [13-14]. Bayesian network is specified by two elements:

A graphical component composed of a directed acyclic graph (DAG) where vertices represent events and edges are relations between events.A numerical component consisting in a quantification of different links in the DAG by a conditional probability distribution of each node in the context of its parents.

Naive Bayes is a very simple Bayesian network, which is the root from the same DAG node (called mother) that node is ignored, and several children, the node corresponding to the observed and the strong assumption of independence between in the context of their parent child nodes.

The classification is by considering the parent node is a hidden variable, that of which should be set for each test object class and sub-nodes represent different attributes of the specified object.

Thus, in the presence of a training set, we have to calculate the conditional probability, because the structure is unique. Once the network is quantified, it is possible to provide the classification of any new object of their properties' value using the Gulf rules. Expressed as:

$$P(c_i | A) = \frac{P(A | c_i).P(c_i)}{P(A)}$$

Where ci is a possible value in the session class and A is the total evidence on attributes nodes. The evidence Acan be dispatched into pieces of evidence, say a1, a2... anrelative to the attributes A1, A2,...,An, respectively. Since naive Bayes work under the assumption that theseattributes are independent (giving the parent node C), their combinedProbability is obtained as follows:

$$P(c_i \mid A) = \frac{P(a_1 \mid c_i).P(a_2 \mid c_i),...,P(a_n \mid c_i).P(c_i)}{P(A)}$$

Note that there is no need to explicitly compute the denominate or P (A) since it is determined by the normalization condition.

C.NBTree

Algorithm NBTree decision tree is a cross between Naive Bayes classifier and classification. NBTree model Best described as a decision tree with nodes and branches Leaf nodes of the Bayesian classification. As with other tree-based Classification, NBTree through, with branches and nodes. Given the A set of instances of the algorithm evaluation node "Practice" for each division of the property. If the greatest value the property is significantly better than the practices Instance, based on the current node, will be divided into Property. If you do not divide, providing an important to better the effectiveness of a naive Bayesian classifier to create the current node. The effectiveness of compute nodes discrete data and five times the cross-validation performed Using Bayesian estimation accuracy [15].

D. Naive Bayes (CFSGSW)

Naïve Bayes algorithm has already discusses in above step and hears discuss the CFSGSW techniques.
According to R.Kohavi and G.H.John (1997) feature selection keeps the original features as such and selects subset of features that predicts the target class variable with maximum classification accuracy [16]. M.Hall (1999) proposed the Correlation Feature Selection measure (CFS) which computes heuristic measure of the “merit” of a feature subset from pair-wise feature correlations and a formula adapted from test theory [17]. Mark A. Hall, Lloyd A. Smith (1998) compared CFS with wrapper approach. The results showed that CFS is comparable or better than Wrapper [18]. Zhiwei Ni et al., (2007) proposed CFS method based on Genetic algorithm to select the optimal subset of attributes. The proposed method was capable of identifying the most related subset for classification and prediction but decreasing the classification precision [19].

The CFS measure evaluates the subset of features based on the two concepts: feature-feature correlation and feature classification correlation. The feature-feature correlation indicates the correlation between two features, while feature classification correlation says how much a feature is correlated to a specific class.

F. Improved Self Adaptive Naive Bayesian Tree

Naïve Bayesian Tree algorithm that has already discusses in above step and hears discuss the Improved Self Adaptive Naive Bayesian Tree. In a given training data, D = {A1, A2,…,An} of attributes, where each attribute Ai = {Ai1, Ai2,…,Aik} contains attribute values and a set of classes C = {C1, C2,…,Cn}, where each class Cj = {Cj1, Cj2,…,Cjk} has some values. Each example in the training data contains weight, W = {W1, W2…, Wn}. Initially, all the weights for examples of training data have equal unit value that set to Wi = 1/n. Where n is the total number of the training examples. Estimates the prior probability P (Cj) for each class by summing the weights and how often each class occurs in the training data. For each attribute, Ai, the number of occurrences of each attribute value Aij can be counted by summing the weights to determine P (Aij). Similarly, the conditional probability P (Aij | Cj) can be estimated by summing the weights how often each attribute value occurs in the class Cj in the training data. The conditional probabilities P (Aij | Cj) are estimated for all values of attributes. The algorithm then uses the prior and conditional probabilities to update the weights. This is done by multiplying the probabilities of the different attribute values from the examples. Suppose thetraining example ei has independent attribute values {Ai1,Ai2,..., Aip}. We already know the prior probabilities P (Cj) and conditional probabilities P (Aik | Cj), for each class Cj andattribute Aik. We then estimate P (ei | Cj) by

$$P(ei \mid C_j) = P(C_j)\pi k = 1 \rightarrow pP(A_{ij} \mid C_j)$$

To update the weight of training example ei, we can estimate the likelihood of ei for each class. The probability thatei is in a class is the product of the conditional probabilities foreach attributes value. The posterior probability P(Cj | ei) is thenfound for each class. Then the weight of the example isupdated with the highest posterior probability for that example and also the class value is updated according to the highest posterior probability. Now, for each attribute Ai, evaluate the utility, u (Ai), of a spilt on attribute Ai. Let j = argmaxi (ui), i.e., the attribute with the highest utility. If uj is not significantly better than the utility of the current node, create a NB classifier for the current node. Partition the training data D according tothe test on attribute Ai. If Ai is continuous, a threshold split issued; if Ai is discrete, a multi-way split is made for all possible values. For each child, call the algorithm recursively on the portion of D that matches the test leading to the child. The main procedure of proposed improved self-adaptive naive Bayesian algorithm is described [20].

TABLE.1
PROS AND CONS: INTRUSION DETECTION ALGORITHM

Algorithm	Pros	Cons
Best First Tree	Creates simpler, less complicated trees for some datasets	A newer techniqueevaluated for subset of decision tree situations
C4.5 Tree	Classifies data with missing attributes	Slower to classify than other techniques.
ID3	Extensively tested a-nd used across many Different application	Unable to handle both discrete and continuous variables.
Naive Bayes	Naive Bayes tend to be attack specific and build a decision tree based on Special characteristics of individual attacks.	The result of Bayesian inference depends strongly on Prior probabilities.
NB Tree (ISA)	Analyses the large volume of network data and considers the complex properties of attack behaviours to scaling up detection rates and reducing false positives in	Not use in real computer network.
NB Tree	NBTree spans out with branches and nodes. Givena node with a set of instances the algorithm evaluates the 'utility' of a split for each attribute.	high time complexity

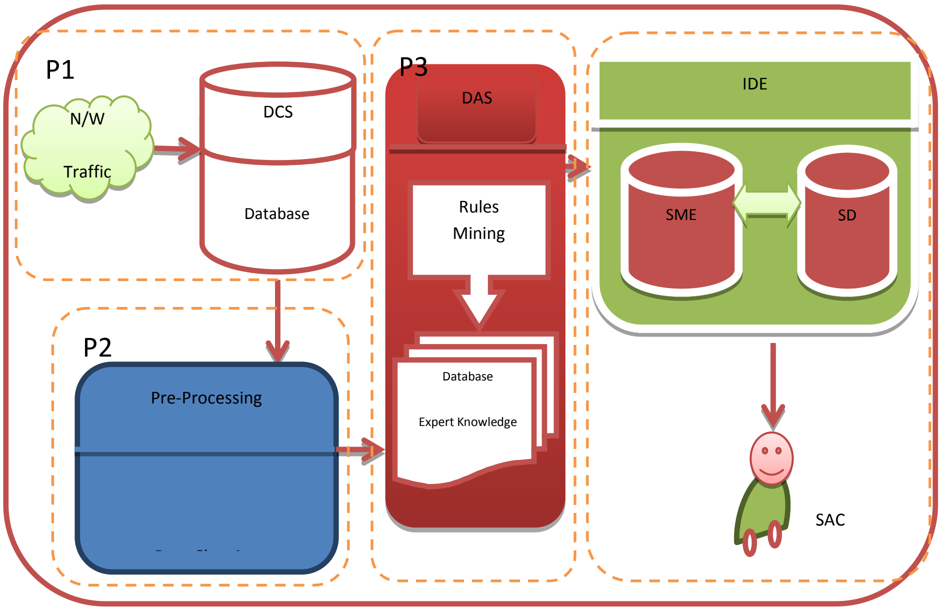
In Table.1 We discuss the ponce and cons of the different Data mining Classification Algorithm. With the help of this table we can identify the Advantages of the particular algorithm and weakness of the algorithm and it is useful to develop a new intrusion detection algorithm.

IV. FRAMEWORK OF INTRUSION DETECTION SYSTEM

Here we had introduced a new framework that use for the intrusion detection in the area of data mining. The proposed framework has divide in to four major phases, in phase I we cover the DCS: Data capturing software, in phase II wediscuss the Pre-processing, in phase III we declare the Data Analysis System which contain the data mining rules which use the expert knowledge while in phase IV contain the IDS and SAC. The detail description of the above phase is discuss below

A. Phase I:DCS

The DCS is collector of traffic. It is simply an interface of capturing information flowing by the system/network interface which is the incoming of the network traffic. When packet capture comes, the system decodes the data in the form of the TCP/IP protocol. The information gathering at DCS with help of some open source network-base intrusion detection system/software. This will collect all the data (which contain noise) and store in Database. Database contains all the data which is basically some row data. The row data contain the network information like the port number, source network address, destination port address and many more.



1) DCS: Data capturing software
2) DAS: Data Analysis System
3) SME: Signature Match Engine
4) SD: signature Dataset
5) SAC: Security Analysis Center / Administrator

B. Phase II: Pre-processing

In Pre-processing Data come from DCS which is basically a row data which contain the missing data, duplicate data, erroneous data or heterogeneities. Which give the birth to data cleaning process. The Pre-processor handles the raw network packets from Raw Traffic Database into a format data that the Data Mining Engine can utilize. In the process, all the erroneous and repeat data packets will be removed by this module. In additional, this module will perform the noise elimination. Generally, the data cleaning face convert the raw data into the form of server tuples <HostID, UsrID, Resource, Action, Time, Duration, ErrorCode>, where presents the host's identity, user's identity, resource name, the action to the accessed resource, the start time of action, the time occupying the resource, and the returned error code. Now this clean data is forwarded to the phase III (Data Analysis System).

C. Phase III: Data Analysis System

The Data Analysis System contains the rule mining and expert knowledge. Hear with the help of database expert knowledge it create the data mining rules such as classification rule, association rule mining, and sequence rule mining. The Data Analysis System also is the core component of the system. Having receiving analyseddata from pro-processor, the module compares thedata with those data read from rules database. Thecompared results are forward to the phase IV.

D. Phase IV: Intrusion Detection Engine.

IDE contain two components SME: signature matching engine and SD: signature data base. Here the rules base data is check whether it contain intrusion or not. The signature data base contains the signature of the different types of attack. The rule bases data which comes from the phase III is served to The Signature matching engine and match the signature in the original data base with the help of signature data base .If SME found any signature match then it alarm to the Security Analysis Center / Administrator which is the end of the intrusion detection cycle.

V. Future Work and Conclusion

In future work we have plan to implement the proposed framework in the real network environment. We would like to test the discussed algorithm to the propose framework and find out the successive ration of the framework in the real platform of the network.

We have present the comparative review on the different algorithm Like Decision Tree, Naive Bayes, Naive Bayes (CFSGSW), NBTree for intrusion detection. We have find the Pros and Cons for each. Naive Bayes tend to build a decision tree based on Special characteristics of individual attacks where in opposite the result of Bayesian inference depends strongly on Prior probabilities. A NB Tree (ISA) analyses the large volume of network data and considers the complex properties of attack behaviours where in opposite it is not use in real computer network and in simple NB Tree it spans out with branches and nodes.

References

[1] D. E. Denning, “An Intrusion-Detection Model”, IEEE Transa-ctions on Software Engineering, vol. SE-13,no. 2,pp.222-232, 1987.

[2] Lazarevic, A., Ertoz, L., Kumar, V., Ozgur,. A., Srivastava, and J., “A comparative study of anomaly detection schemes in network intrusion detection,” In Proc. of the SIAM Conference on Data Mining, 2003.

[3] T.H. Ptacek and T.N. Newsham. lnsertion, Evasion and Denial of Service: Eluding Network lntrusion Detection. Technical report, Secure Networks, January 1998.

[4] J.P. Anderson. Computer Security Threat Monitoring and Surveillance. Technical report, James P. Anderson Co., Fort Washington, PA, April 1980.

[5] PengLiu et al .The Design and Implementation of a Self-Healing Database System Database System. School of info Sciences and Technology Department of Information Systems, Pennsylvania State University UMBC, University Park, PA 16802 Baltimore,MD 21250.

[6] Pramote Luenam, Peng Liu, .ODAM An On-the-tly Damage Assessment and Repair System for Commercial Database Applications., Dept. of Info. Systems, UMBC Baltimore, MD21250.

[7] Fayyad, D, Piatetsky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery in Databases. Al magazine, 17(3):37.1 54.1 996.

[8] Han, J. and Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publisher.2000

[9] Daniel T, Larase, Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons, Inc, 2005.

[10] Quinlan, J. R.: C4.5, Programs for machine learning.Morgan Kaufmann San Mateo Ca, 1993.

[11] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C.J.: Classification and regression trees. Monterey, CA Wadsworth & Brooks, 1984.

[12] Quinlan, J. R.: C4.5, Programs for machine learning. Morgan Kaufmann San Mateo Ca, 1993.

[13] Jensen, F. V.: Introduction to Bayesien networks. UCL Press, 1996.

[14] Pearl J.: Probabilistic Reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmman , Los Altos, CA, 1988.

[15] R. Kohavi, “Scaling Up the Accuracy of Naïve-Bayes Classifiers: A Decision-tree Hybrid”, Proc. of the 2th International Conference on Knowledge Discovery and Data Mining, pp.202-207, 1996.

[16] R.Kohavi, G.H.John, “Wrappers for feature subset selection”, Artificial Intelligence, vol.1, no.2, pp.273-324, 1997

[17] M. Hall, Correlation Based Feature Selection for Machine Learning, Doctoral Dissertation, University of Waikato, Department of Computer Science, 1999.

[18] Mark A. Hall, Lloyd A. Smith, “Feature Selection for Machine Learning: Comparing a Correlation based Filterapproach to the Wrapper”, American association for Artificial Intelligence, 1998.

[19] Zhiwei Ni, Fenggang Li, Shanling Yang, Xiao Liu,Weili Zhang, QinLuo, “Attributes Reduction based on GA-CFS method”, Proc. of the joint 9th Asia-Pacific web and 8th Intl. Conf on Web-age information management on Advances in data & web management, Springer-Verlag,2007.

[20] Dewan Md. Farid, Nguyen Huu Hoa, Jerome Darmont, Nouria Harbi, and Mohammad Zahidur Rahman “Scaling up Detection Rates and Reducing False Positives in Intrusion Detection using NBTree”

BIOGRAPHY

	Hemant H Patel has completed the M.Tech in Information Technology from U.V Patel College of Engineering, Ganpat University, India. His Research area is Intrusion Detection in Data mining Subject. Right now he is work as Assistant Professor at CSE/IT Department in Dr.Subhash Technical Campus, Junagadh, Gujarat-362001, India.
	Bharat.R.Sarkhedi is pursuing in M.Tech in Computer Engineering from Patel collage of science &Technology, RGPV University, Bhopal, India. His Research area is Intrusion Detection in Data mining Subject.
	Hiren.M.Vaghamshi is pursuing in M.Tech in Computer Engineering from Noble Engineering College, GTU, India. His Research area is Intrusion Detection in Data mining Subject.