Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345 (Online)
An Open Access, Online International Journal Available at http:// http://www.cibtech.org/sp.ed/jls/2014/01/jls.htm
2014 Vol. 4 (S1) April-June, pp. 417-425/Ogbi and Hazbavi
Research Article

EVALUATION OF DATA MINING IN INTRUSION DETECTION SYSTEMS

Masoud Ogbi[1] and *Jasem Hazbavi[2]
[1] MS Student in Computer Engineering, Science and Research Branch, Islamic Azad University, Khuzestan, Iran
[2] Student in Computer Engineering, Institute For Higher Education ACECR Khuzestan (Ahwaz), Iran
**Author for Correspondence*

ABSTRACT
Data mining One of recent advances in order to IT data manage. Data mining is a set of techniques that allows one to move beyond conventional data processing and to help for information derive which secret or hidden in massive data. In fact, data mining to find models and data finders are experts who used a special software for find data systematically and without regulation of large sets of data. In this article we will discuss the applications of data mining in intrusion detection systems.

***Keywords**: Data Mining, Data Processing, Application Systems, Intrusion Detection*

INTRODUCTION
Intrusion Detection before being raised to data mining
Firewalls play a crucial role in securing the network safety. They are as much as possible to prevent the offender intrusion to the network, but also the strongest firewalls cannot make entirely secure and some attackers can get some of them to carry out acts of sabotage. Fortunately only in the firewall logs can be found signs of the attack, which after it occurred. For example to further clarify the issue, consider the operation of an Internet worm. Within the network, a user will receive a mail from his friend, for example, the latest versions of his favorite games. Running a Trojan horse program is installed on the user's system when the user does not have work and leisure, to open a connection to the server for it to start sending important information within the network. Such a mechanism cannot be identified by filters because these servers often run on a virtual ports such as HTTP. In this case, only the proxy can do even better. They to run owns server on port FTP (or any other port which through binary information is exchanged). Server and Trojan horse program are adjusted before sending the original information to exchanged a series of packets, pretend that they can establish a virtual connection. Such an attack by proxy is not detectable. In this case, even in the firewall log is not seen as a sign of attack, because all transactions were allowed. This section to introduces the invasion finder or IDS systems to examine how the protects the network against attacks, the firewall are not able to recognition them.
Raid finder systems are software that examining network traffic and off a series signs to recognize these attacks with. These types of systems are used along with firewalls and to make additional security. An invasion finder system can be just to protect from a system or all systems on the network the latter called NIDS. Invasion finder systems treatment against suspected case and their attacks, Divided into active and passive categories. Active invasion finder systems can be schedule so that as soon as occurrence of a suspected case, to indicate the appropriate response (suspected disconnection generate an alert...), but only active passive invasion detection systems are recorded the events which can then be investigated. There are two different techniques for attack detection systems that following to be cited.

Invasive Detection Systems Based On Inspection:
These types of systems are the most common type of invasion detection. How they function so that monitor a system or network activity and through a set of rules defined or comparison by a normal state to discover the suspected cases. Invasive finder are used to explore of suspected cases the following factors:

• Signs that are obtained from the analysis of network traffic, such as Port scan a connection to the unauthorized port
• how the utilization of system resources such as CPU and memory or the communication network at the sign of in unusual times which is automatic sentence.
• how to use the system file, such as newly created files, file system changes with changes in the users account.

Invasions Finder Systems Are Several Different Types:
• legality Invasion finders: The systems can identify attacks from a set of rules. Each rules specifies a known method to attack which to use a series of symbols or a sequence of activities to discover an attack. Connect to a specific port, or a changes in operating system files, including are the symptoms.
• Statistical Invasion finder : the system to discover the attacks with comparison of the system current state to its natural state. This method is suitable for detecting new and unknown attacks, but instead is able to detect many known attacks are not very easy to take them as normal behavior of system. It is clear that this method can detect attacks once they occur.
• hybrid Invasion finder : These type of systems to utilize the combines of two previous techniques. detect Known attacks by appropriate legislation and unknown attacks by statistical analysis. Many of these systems are capable of detecting an unknown attack, as well as to make laws relating to it, that since used to identify the type of attack used.
• invasion Finder System of prey: these type of systems to appear as a service over the network But as soon as someone is going to use them to generate a security warning, A raid finder bait looks like a server that is not well protected and therefore can attention quickly. Attracted the attention of attackers through is attracted this system that for network does not have any functional significance Upon first use, it is invasive and a security alarm is generated to notify the network administrator of the attack. Invasive detection of prey can be one or more servers on a single server, a single server or network, single (rarely). In the first case, the invasion finder is create the multi-service TCP / IP on a system that from them is not used in the network. (e g Service charge echo or NFS) In some cases, these services have not even the service performance and application response only returns timeout and only will recognize the connection to the port. Effective solution is used a full system as bait. The only thing to be done is to A copy of the operating system along with all the usual services create on Prey System And NAT network can be configured to send all requests to the system. Then add the necessary rules to specific services such as WWW and SMTP traffic to the real servers of the network combines these looks and sees that they all belong to a single system. Since the attack to prey services (NFS and NETBIOS) is far easier, attackers prefer to try them on first, and this will generate a security alert. At this stage, the offenders can be permitted to carry out their work and identified the methods that they use to attack the system. After some time the offenders will understand that proper security measures by the network administrator to deal with them is considered. one can even put a copy of the network services on the baits system main the assailants later found to be subject and administrators have a greater opportunity to secure servers. In project the intrusion detection of certain parts of data mining which is used as follows:
• Deleting normal data from suspected attack data will allow analysts that the gives more focus to finding a real one .
• Producers recognize false declarations
• Find unusual activity can be detected in real attack .
• identify patterns , detect IP addresses and similar activities

For these activities , the following methods are used by data finder :
• Summarize data by statistics , find out-of- range values
• Detection : provides a graphical summary of data
• Cluster data in natural categories

• The discovery of association rules, the definition of normal activities and discover unusual cases

DATA MINING AND INTRUSION DETECTION

data mining One of recent advances in order to IT data manage. Data mining is a set of techniques that allows one to move beyond conventional data processing and to help for information derive which secret or hidden in massive data. In fact, data mining to find models and data finders are experts who used a special software for find data systematically and without regulation of large sets of data. In project the intrusion detection of certain parts of data mining which is used as follows :

• Deleting normal data from suspected attack data will allow analysts that the gives more focus to finding a real one.
• Producers recognize false declarations
• Find unusual activity can be detected in real attack.
• identify patterns, detect IP addresses and similar activities

For these activities, the following methods are used by data finder:

• Summarize data by statistics , find out-of- range values
• Detection : provides a graphical summary of data
• Cluster data in natural categories
• The discovery of association rules enabled the definition of normal activities and detect unusual cases
• categories : predict categories that encompass specific records.

Start with a realistic look

• A misleading fact is that automation equipment can solve all the problems of human conflict makes it unnecessary. It's not a mirage intrusion detection. Human analysts for observe optimal performance or non-performance mechanical systems, identify new classes of attack craft, always will be needed. The need for intrusion detection analysts are increasing every day.
• In some automated intrusion detection system's ability to respond promptly is very important and necessary.

Establish appropriate

• a suitable platform for intrusion detection using data mining process requires the following cases:
• Database: Why We Need a local, specific structures and data are stored in standard data database should be updated regularly and should have a rapid response mechanism to be research. For this purpose, should be used as database management systems.
• Workspace: To use an intrusion detection system we need to workspace and data that are well suited for data mining also should be involved in data computing and metadata storage as fast as copy the normal data. The workspace should have a diverse set of sample data for conduct data mining experiments. Also the storage devices are required for a specific application.
• Ability to calculate: data mining processes by data mining tools and accessories to the CPU and main memory needs a lot of computing power and the increased CPU speed and memory capacity is normal. Studies have shown that intrusion detection using data mining need to four times more resources than doing intrusion detection without use of data mining.
• software: In addition to having the base software such as operating systems, database functionality and high quality standard, require to data mining tools along with the right the possession of their property. Many tools have been developed for data mining processes that can be noted to Clementine, eka , Gritbot , It is merely a data mining tool for data mining process is insufficient and requires expertise and personalized tools that can be effectively used and have to relevant work experience.

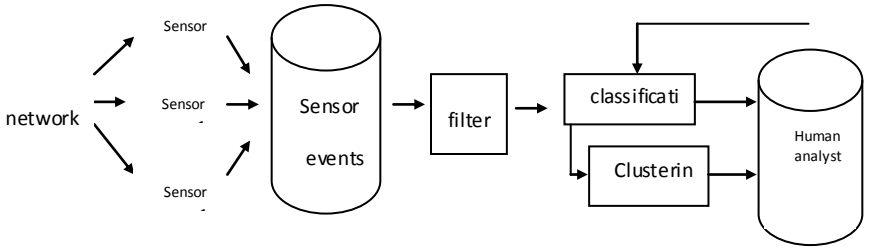
DESIGN, CALCULATE AND STORE THE APPROPRIATE OPTIONS

• Data records have a large number of attributes. When we use data mining to intrusion detection in computer network, we must use from the TCPDUMP level data:

One is Sniffer class or to simply sentence actually is a network traffic analysis. Software is an old and famous TCPDUMP based on Unix operating system family. Actually is a network traffic analyzer. An analysis of network traffic, commonly called remember Sniffer, it is the duty of examining packets exchanged on the network is responsible for Using a Sniffer, specifying a particular network interface, Can by monitoring and analyzing network packets exchanged on the network interface is attached to it. In other words, a Sniffer can be likened to a monitoring system which to store and review all information transmitted on a physical medium or use data of alert level. The fields for each data type will have source and destination IP address, source and destination port numbers, date, time transport protocol (TCP, UDP, ICMP, ETC) and the traffic start and end time. The basic characteristics, to give The proper description of a particular relationship or give a warning, but often not sufficient to identify unconventional and suspicious.

DATA FILTERS INSTALLATION

• Sensors in the event table represents over 95% of traffic is mapped to IP addresses. So that a source IP tries to communicated with hundreds or thousands more destination. Security professionals can input data in a data mining process can be conditioned to filter network traffic.



For Example, Figure 1, In Network Intrusion Detection

Refining The Overall Architecture For Intrusion Detection:

The figure above shows an architectural model for intrusion detection. Network traffic is collected by different sensors. The sensors periodically to collect and send for transmit data to a central server-based database. Events sensed by sensors before it reaches the classifier and clustering are filtered. Data mining tools to be filtered false positive errors and unusual behavior among the remaining data. The purpose of this operational model having all the warnings by human review analyst.

Classification Rules:

Classification of samples are used in order to allocate the predefined categories. Machine learning softwares expressed , this act by extracting and learning accurate classification rules for dividing the sample data. Classification models can be created using different algorithms. Henery [1994] algorithms are classified into three types:

•Share this line (such as sorting logistics and ...)

•Decision-tree and rule-based methods (algorithms, CART, AQ, C4.5, etc.)

•estimates density of poignancy (KNN, NB)

•among mentioned classification of decision tree method, be used two methods based on the law than other methods, Because The results output by this method is more understandable to humans.

• Examples of good: training data quality, one of the most important factors is a high-performance classifier. The quality of the training data as a function of the number of samples to be representative of how records and features is used to describe it.

• labeled data: classification with the help of labeled samples are used. Data usually are labeled by the expert humans. In this research is used from labeled data for category.

METHODS

In This Study

In general, expression characteristics and behavior of computer network attack and penetration participants is usually very difficult and requires an expert. In addition, with the advancement in computer networks, number of attacks and intrusion are more and more. In fact, knowledge that comes with expert humans after time, loses its value And should be updated and the system will be placed on, the same factors always feel the need to be an expert. In machine learning techniques to extract knowledge from the data itself and the same factor has faded the role an expert. from steps are used to check in the algorithm machine learning techniques.

THE DATA SET

In this step, the data set is determined. Usually the data are analyzed by the intrusion detection through sensors and computer networks are available. In this step, the data reviewed in detail and determined sincerity or insincerity of the data. Missing values is determined by using different placement method and format of the data. In Most cases, the data in terms of scale, not a single order using normalization methods, become as unit.

DATA TECHNICAL COMMENT BY THE MODEL ANALYST

In this step, using expert knowledge and information, through calculating information such as weighted, average, and a data center. . . Analysis is performed on the data. Usually this is done in order to have an overview of the data and to create a model of the data.

TRAINING MODEL

After creating a model that can be taught to it the data that are used to teach the training dataset is referred, that including the predictor variables records and the values categories for the variable target. In fact, a models is created from the data contained in this dataset will do learning practice.

KNOWLEDGE CREATION

After teaching, the model created have the knowledge that it has learned from the training data set. The knowledge include the data structure and knows the patterns in it.

TESTING THE MODEL

After gaining to knowledge from the training data set, can test this knowledge to datasets that do not provide any information about them.

USE OF MODEL

After testing the model, it can be used in real environments and its outcome and results can be used in making decisions.

SELECT THE DATA SETS

In 1998, in order to evaluate intrusion detection, program under title of 1998 DARPA by MIT Lincoln Laboratory was created, that aimed to evaluate the intrusion detection. To this end, will be create a standard data set that have different attacks simulated in a real network environment. Intrusion detection contest KDD99, the data set used. The dataset used in this research is KDD99. This data set is widely accepted and used as a standard for the evaluation of intrusion detection systems. This data set contains both the training and testing of data. The data by MIT Lincoln Laboratory were collected, which has a network of machines of different types of prey. The bait machine number in outside and inside of network was used for data collection. And the machines inside and outside of network, were responsible the

created networks traffic. And by whom, They also attacked by the background this traffic to a victim machine. The following figure, illustrate a scheme designed from the network to collect data KDD99.

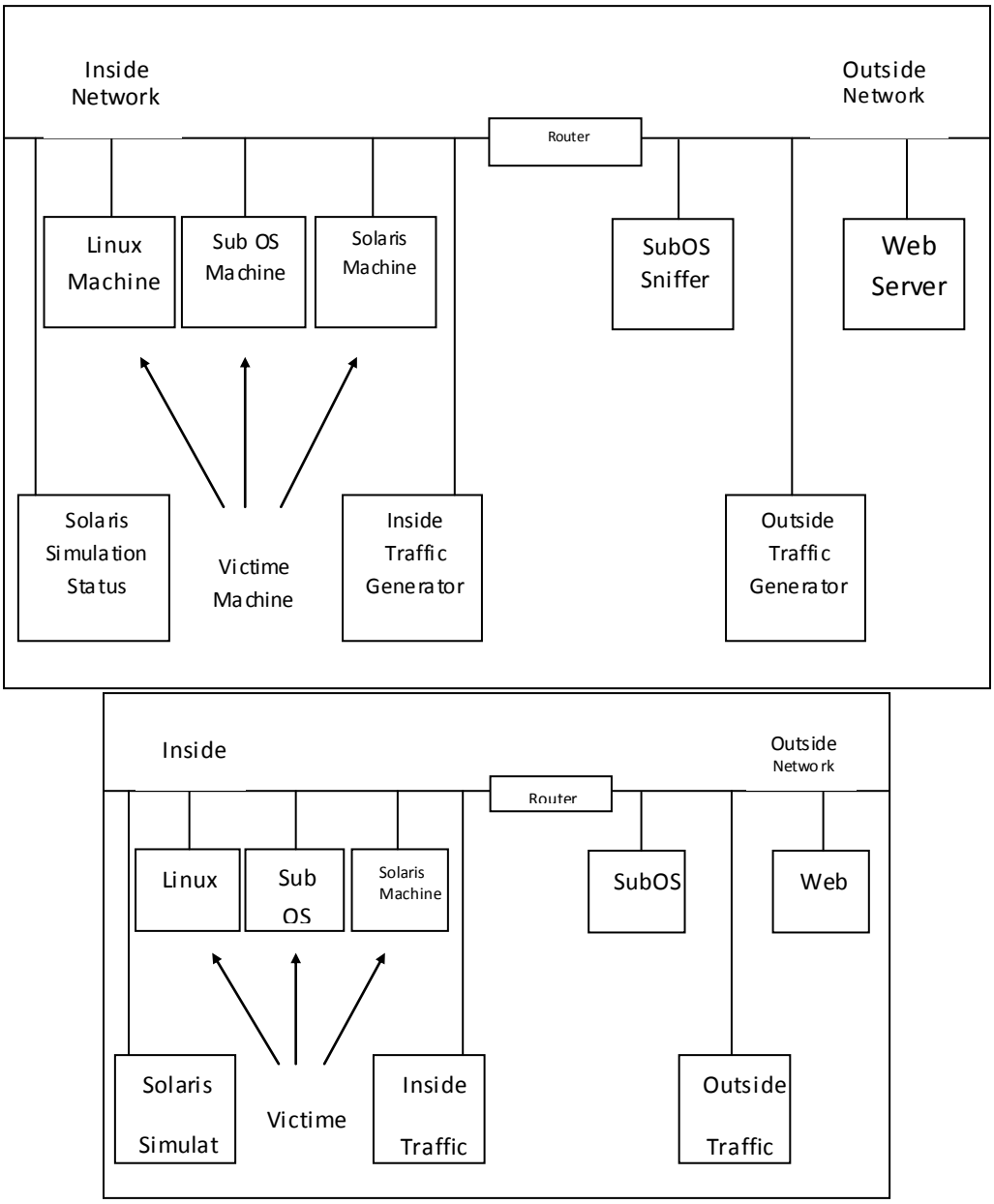


Figure 2: Scheme Of The Network Was Designed

To Collect Data KDD99

In this study, the percentage of the total training data was used. Any record in This data is a normal record or belong to one of 22 different categories of attacks. All these attacks are divided into four main categories which are:

DOS: In this attack, the system resources are used too much and causes normal calls requests to rejected for the availability of resources.

R2L: in R2L type attack, the remote attacker with remote unauthorized intrusion into the victim's car, began to abuse of the user statutory accounts and attempts to send packets on the network. U2R: This type of attack from victim's car to runs successfully and takes root. Probing: In this type of attack computers are scanning to gather information or find known vulnerability capabilities.
Examples of these four separate attacks in Table 3 are listed.

Table 3: Examples Of Common Assault

DOS (Denial Of Erivce)	R2L (Remote-To-Local)	U2R (User-To-Root)	Probing
Ping Flood Syn flood Mail Bomb	Dictionary FTP Write Send mail	Perl Fol format-local module Eject	IP Sweep Saint satan
PDOS			

KDD99 Dataset Features:
Dataset KDD99 41 is a feature. Some of these features are continuous and some discrete. Number of attributes used to evaluate the data mining techniques are different on data sets, sometimes all these features and some subset of these features are used. The full list of features for communication records in Table2are listed. The pattern are available by machine-readable data. The features are enumerated to three main categories, content and traffic, each of which is described in a separate table. Among the features included in this data set have Only three categories of attributes, Flag and Protocol Type and Service Type that should be placed numeric values instead of numeric values . One of the methods that can be used for this purpose that is in all of data frequency , each of the values are computed three features or to be used instead of three features and or can not use in algorithm
(Kind of features is shown with the letters S (symbolic) and C (constant))

Table 4: Table Of Attributes Names KDD99

	Calculated value		
		normal	attack
Actual value	Normal	A	B
	Attack	C	D

NORMALIZATION

The data contained in these datasets have different values so that the maximum and minimum values are very different to another. If reject normalization between data, the attributes that have large amounts will dominate character with small amounts of feature and will be unreasonable results. Therefore, it is essential that this normalization to be used, for this purpose to applied from normalization max - min ,wherein Newval represents the new value, val current value of data and the min and max represent the minimum and maximum values.

Newville = normalize (In (vale + 1)) (1)

$$\text{Normalize (Xi)} = \frac{\text{Xi - mini}}{\text{Maxi} - \text{mini}}$$

CREATE A SET OF TRAINING DATA

In data mining, by machine learning methods with the guide and without it should already have enough information about the data, It can be created by use of tagged records, specific training data set with a certain percentage of normal and attack records created and used in the evaluation. Since the algorithms

behave differently than the training data set, to percentage of normal and different attack records, it created the data which foregone percent of each record to be seems necessary and useful. Therefore, in this project the training data with normal percentages of attack have been studied.

SELECT A METHOD FOR LABELING CLUSTERS

Output performance of clustering algorithms, which are usually given by the clusters are separated. Here should be the common practice in this area, the clusters can be labeled. Followed by a transition labeled data in clusters, each cluster can be studied. And algorithm performance measures can be calculated and for these criteria, the values obtained to compare the algorithms together are place criteria .from Conventional methods can be used to label the cases include:

The method is based on the count (Count-Based): In this method, the cluster have a small number of records as tagged attacks and other include normal records, these methods are common practices in cluster labeling s.

The method is based on the distance (Distance-Based): In this method, the cluster that have been separated is high from other clusters and their distances from the other clusters, are considered as abnormal or invasion cluster and adjacent clusters, including data are normalized.

SELECTION CRITERIA FOR PERFORMANCE EVALUATION

to evaluating and compare the performance of algorithms to be used from Different criteria. These criteria are detection rate (DR), false positive rate (FPR), recall , accuracy and F-measure that after form the scattering matrix for the clusters are easily calculated. Scattering matrix is shown in Figure 4-3. Where 9% of the normal data in the normal clusters, b is percent of normal data in clusters of attack, C is percent of attack data in normal clusters and d is percent of attack data in attack clusters. Amounts DR, FBR, Recall, Precision, Accuracy, F-measure for this matrix are calculated by formula 3-2 to 3-6.

$$DR = \frac{d}{d+c} \quad (2)$$

$$FPR = \frac{b}{a+b} \quad (3)$$

$$RECALL_{NORMAL} = A / (A+B) \quad (4)$$

$$RECALL_{ATTACKTYPE} = B_{ATTACKTYPE} / (A + B_{ATTACKTYPE})$$

$$PRECISION_{NORMAL} = A / (A+C) \quad (5)$$

$$PRECISION_{ATTACKTYPE} = B_{ATTACKTYPE} / (BATTACKTYPE + D) \quad (6)$$

$$F\text{-}MESURE = \frac{2 \times Recall \times Precision}{Recall + Precison}$$

Figure 5: scattering matrix of clusters

Figure 5: Scattering Matrix Of Clusters

Top value is equal to value a and fn value is equal to b, Fop value is equal to the value of c and ten amount is equal to d.

APPRECIATION AND THANKS

The National Iranian oil products distribution company that helps me out in this matter thanks in advance and I appreciate.

REFERENCES

H. Miller, and j. Han, geographic data mining and knowledge discovery taylor and francis، .london، u.k، 2001.

M. Steinbach, p. Tan, v. Kumar, s. K, and c. Potter, data mining for the discovery of ocean climate indices, proceedings of the 5th workshop on scientific data mining (sdm 2002), (arlington, va, apr. 13), society of industrial and applied mathematics, pp 7–16, 2002.

S j . Stolfo , w. Lee, p. K. Chan, w. Fan and e. Skin, "data mining-based intrusion detectors: an overview of the columbia ids project", sigmod record, vol. 30, no. 4, december 200 pp 5-14, 2000.

H. Ka rgupta, a. Joshi, k. Siva kumar and y. Yes ha, "data mining: next generation challenge sand future directions", prentice hall of india, pp. 157- 219 2005**.**

L . A.f. park‹ k. Ramamohanarao‹ and m . Palaniswami, "fourier domain scoring: a novel document ranking method", ieee transactions on knowledge and data engineering, vol. 16, no. 5 , pp529-539, may 2004.

S. Schockaert, m. De cock, c. Cornelis and e. E. Kerre "efficient clustering with fuzzy ants", applied computational intelligence, world scientific, p. 195-200, 2004.