International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 4, Issue 5, May 2015 **ISSN 2319 - 4847**

Intrusion Detection System Using Means of Data Mining By Using C 4.5 Algorithm

[1]Meghana Solanki , Vidya Dhamdhere[2]

[1]Post Graduate Student, G.H.R.C.E.M, Wagholi, Pune, Maharashtra, India

[2]AssistantProfessor, G.H.R.C.E.M, Wagholi, Pune, Maharashtra, India

ABSTRACT

Intrusion compromise Security as well as privacy of a system. Intrusion Detection System (IDS) plays fundamental role in network security as it detects numerous categories of attacks in case of network. For this reason, we are proposing Intrusion Detection System using data mining technique: C 4.5 .Here, Classification will be done by using C 4.5 and some experiments using KDD Cup'99 dataset are conducted for confirming the effective in case of proposed system . The C 4.5 is used as the most famous classification algorithms in case of data mining area. For implementing proposed system KDD Cup'99 data set is used to perform some experiment. from results it is proved that time required to build C 4.5 model can be reduced by using proper data set pre-processing. Also system increases attack detection rate in case of C 4.5 .it also decreases False Positive Rate (FPR) .

Keywords: Classification, Intrusion Detection System (IDS), Kernel Function, KDD, Pre-processing, C 4.5 etc.

1. INTRODUCTION

An intrusion detection system (IDS) checks out all inbound and outbound network traffic and identifies suspicious patterns. It may indicate a network or system attack. It also indicates that someone attempting to break into or compromise a system. An ID means Intrusion Detection acts as a security management system for computers as well as networks. An ID system collects information. It analyzes information from various areas in a computer field to identify possible security crack. User behavior is compared with user profile to detect unauthorized users in the system. An unauthorized entry into the computer system is indicated by the event, such event is detected by the system. It notifies not only a control function about the unauthorized users but also events which show unauthorized entry into the computer network. Also it contains a control function which mechanically takes action in response to the event. There is a dynamic creation of user profiles are for each computer user when the computer user logs into the computer system first time. The user's profile is dynamically updated after subsequent logins. False alarms are reduced by performing comparison of user behavior with the dynamically built user profile. The system also includes different types of functions such as log monitoring function, a port scan investigator and a session detector function. An intrusion detector has sensor for monitoring a condition and another sensor for monitoring different condition in a space to be protected against intrusion. An Intrusion Detection System (IDS) can be installed in different network processing devices which are distributed throughout a network. An ID performs the role of security management system not only for computers but also for networks. ID uses *vulnerability assessment* also called as *scanning*. It is one kind of technology used to evaluate the security of a computer system as well as network. There are number of Intrusion detection functions available such as observing and inspecting both user and system activities, scrutinizing system configurations and vulnerabilities, examining system and file integrity, Ability to accept patterns typical of attacks, Analysis of uncommon activity patterns, Tracking user strategy violations etc.ID systems are designed for in increasing number of attacks in case of major sites e.g. Pentagon as well as the U.S. Defense Department. The assurance of security is becoming increasingly difficult. The main reason is that achievable methods of attack are becoming ever more complicated. The less technical ability is needed for the novice attacker, because proven old techniques are easily accessed through the Web. Network based intrusion detection try to pinpoint illegal, forbidden, and anomalous behavior depend entirely on network shipment. Host based intrusion detection try to pinpoint unauthorized, illicit, and anomalous behavior on a specific device. Host Based IDS generally includes an agent installed on each system as well as alerting on local OS. it also includes application activity. Physical intrusion detection is the act of recognizing threats in case of physical systems. Physical intrusion detection plays the role of physical controls which is put in place to ensure CIA. In number of cases physical intrusion detection systems also act as prevention systems. There are number of Examples of Physical intrusion detections such as Security Guards, Security Cameras, Access Control Systems, Firewalls, Man Traps, and Motion Sensors. Traditionally, some techniques, such as user authentication, data encryption, and firewalls, are used to protect computer security. Intrusion detection systems (IDS), which use specific analytical technique(s) to detect attacks, identify their sources, and alert network administrators, have recently been developed to monitor attempts to break security [14].

2. LITERATURE SURVEY

In this paper author has presented how Intrusion detection system detects attacks in the network using k-means and C4.5.also it detect anomalies in the network [1].In this paper author has presented various aspects of intrusion detection. Due to which a researcher can become quickly familiar with every aspect of intrusion detection. There are large number of intrusion detection methods, techniques and systems [2].
In this paper author began the concept of intrusion detection. It is started in 1980 [3]; author introduced a threat classification model. This model developed a security monitoring surveillance system. This system is based on detecting anomalies in user behavior. In this paper [4] author has presented one collaborated IDS module. it make a real-time detection as well as block intrusions before occurrences depend on HIDS. It uses sequences of system call for anomaly detection.
In this paper [5] author proposed a system using rough set. It is used for attribution reduction. Support vector machine is used for intrusion detection classification. In this paper [6] author proposed a method of detecting intrusion using incremental SVM. It is based on key r selection.
In this paper [7] author has used RST (Rough Set Theory) and SVM (Support Vector Machine) to detect intrusions. RST is used to preprocess the data as well as reduce the dimensions.SVM model takes the features selected by RST. It is used to learn and test features respectively. In this paper [8] author powerfully has introduced intrusion detection system implemented using Principal Component Analysis (PCA) combined with Support Vector Machines (SVMs) .it is as an approach to select the optimum feature subset.
In this paper [9] author has presented a novel fuzzy class-association rule mining method. This method is based on genetic network programming (GNP) which is used for detecting network intrusions. In this paper [10] author has proposed Dirichlet-based trust management system. It is used to count the level of belief among IDSes according to their mutual experience.
In this paper [11] author has described an incremental network intrusion detection system. It uses a two step architecture. In the first step uses a probabilistic classifier. It is used for detecting potential anomalies in the traffic. In the second step uses a HMM based traffic model. It is used to narrow down the potential attack IP addresses.
In this paper [12] author has proposed Kernelized Support Vector Machine along with Levenberg-Marquardt (LM) Learning. In this paper [13] author has proposed a category-based selection of effective parameters for intrusion detection. Also intrusion detection system combined with Principal Components Analysis (PCA).

3. GENERAL ARCHITECTURE OF OUR APPROACH

3.1. Data Set Collection

To ensure the usefulness as well as the feasibility of the proposed IDS system, we have used 10% KDD 99 dataset.The KDD Cup 99 dataset is most commonly used popular dataset in the field of intrusion detection from the last decade. Many researchers use it for experiment. Many researchers have contributed their work to analyze the dataset. They used different techniques for this purpose. Many industries make the use of analysis which produces as well as consumes data, of course that includes security. The KDD training dataset contains 10% of original dataset. it is approximately 500,000 single connection vectors. Each of which contains 41 features. It is labeled with one particular type of attack such as normal, an attack. Each vector has label that may be either normal or an attack. Drift from ‘normal behavior’ are considered as attacks. Normal attacks contain records with normal behavior. Memory constrained machine learning methods uses a 10% training dataset. The training data set contains 19.69% normal as well as 82% attack connections. KDD CUP 99 is most widely used in attacks in network traffic. The imitate attack falls in one of the following four types:

1. Denial of Service Attack (DOS): In this type the unauthorized person makes a machine or network resource unavailable to the desired users. It temporarily pause services of a host connected to the Internet. It also suspend services for short time.DOS contains the attacks such as 'neptune', 'back', 'smurf', 'pod', 'land', and 'teardrop'.

2. Users to Root Attack (U2R): In this type of attack the unauthorized person logged as a normal end user on the system and is able to exploit some vulnerability which is used to obtain root access to the system. U2R contains the attacks such as 'buffer_overflow', 'loadmodule', ‘rootkit’ and 'perl'.

3. Remote to Local Attack (R2L): In this type of attack the unauthorized person send packets to a remote system over a network without having any account on that system, gain access either as a user or as a root to the system and do harmful operations. R2L contain the attacks such as 'warezclient', ' multihop', ' ftp_write', 'imap', 'guess_passwd', 'warezmaster', 'spy' and 'phf'.

4. Probing Attack (PROBE): In this category the Scans the networks to identify valid IP addresses and to collect information about host. PROBE contains the attacks such as 'portsweep', 'satan', 'nmap', and 'ipsweep'.Simulated attacks are divided into four types as shown in Table I.

Attack Types	Category
Normal	Normal
Apacha2	DOS
Back	
Land	
Netpune	
Pod	
Processtable	
Smurf	
Teardrop	
Udpstorm	
Buffer_overflow	U2R
Httprunnel	
Loadmodule	
Perl	
Ps	
Rootkit	
Sqlattack	
Xterm	
ftp_write	R2L
Guess_write	
Guess_password	
Imap	
Multihop	
Named	
Phf	
Sendmail	
Snmpgetattack	
Snmguess	
Spy	
Warezclient	
Warezmaster	
Work	
Xlock	
Xsnoop	
Lpsweep	Probing
Mscan	
Namp	
Portsweep	
Saint	
satan	

3.2. Data Set Pre-Processing

Pre-processing of original KDD 99 dataset is needed to make it as proper input for C 4.5. Data set pre-processing can be achieved by applying:
i. Data set transformation
ii. Data set normalization
iii. Data set discretization

i) Data set transformation: The training dataset of KDD 99 consist of about 5,000,000 single connection cases. Each connection case includes 41 features including attacks or normal. From these labeled connection cases, we have to convert the nominal features to numeric values in order to make it eligible input for classification using C 4.5. For this transformation, we will use table 2.

Table 2: Transformation Table

TYPE	FEATURE NAME
ATTACK OR NORMAL	NORMAL
ATTACK OR NORMAL	ATTACK
PROTOCOL_TYPE	TCP
	UDP
	ICMPV
FLAG	OTH
	REJ
	RSTO
	RSTOS0
	RSTR
	SF
	SH

An example of original KDD 99 data set record is shown in figure 1.

0 TCP FTP_DATA SF 561 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 33 0 0 0 0 2 0 0 2 6 0 3 6 0.18 0.04 0 0 0 0 0.06 0 NORMAL
0 UDP OTHER SF 257 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 4 2 0 0 0 0 0.09 0.26 0 254 2 0 .7 0.99 0 0 0 0 0 0 0.86 0 NORMAL

Figure 1 The Original KDD 99 dataset

After transformation, the original KDD 99 dataset will become as shown in Figure 2.

0,3,42,25,592, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3,3,0,0,0,0,260,36,0.28,0,0,0,0,0.06,0,0
0,4,27,25,257, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,24,2,0,0,0,0,0.09,0.26,0,244,2,0,0.7,0.99,0,0,0,0,0,0,56

Figure 2 Original KDD 99 dataset after applying transformation

ii) Data set normalization: Dataset normalization is necessary to improve the performance of intrusion detection system when datasets are too large. Using min-max method normalization is performed.

iii) Data set discretization: Dataset discretization technique is used for continuous features selection in case of intrusion detection as well as to create some homogeneity between values, which have different data types. range discretization technique is used for this purpose.

4. INTRUSION DETECTION USING C 4.5

Classification algorithms have attracted appreciable importance both in the machine learning as well as in the data mining research areas. Among classification algorithms, the C4.5 system of Quinlan gains a special mention for several reasons. It represents the result of research in machine learning that traces back to the ID3 system. As such, it often considered as the key point in case of development as well as analysis of novel proposals. from results it is clear that the C4.5 algorithm which is tree induction algorithm gives good accuracy in case of classification .also it is the speedy

algorithm among the compared main-memory algorithms .they are used for machine learning and data mining. It should be mentioned that several external-memory algorithms and parallel implementations have been proposed. The aim is speeding up the execution time and reasoning on very large training sets. In the proposed system, we have developed a C 4.5 model for classification. While intrusion behaviors happen, C 4.5 will find the anomaly. A classification task involves training set as well as testing set which consist of number of cases. Each instance in the training set includes one "target value as well as several "attributes". The aim of C 4.5 is to develop a model which predicts target value of data instance in the testing set. The system design for IDS using SVM is shown in figure 3.
System architecture of Intrusion Detection System is shown in fig.1.System architecture consists of following methods:
K-means Clustering: Clustering is the method of grouping the records into either classes or clusters. Entities within the same cluster have high similarity as compare to entities within the dissimilar cluster. K-means algorithm is useful for undirected knowledge discovery and is relatively simple.
Neuro-Fuzzy: Neural networks are compelling tool for classification. They are flexible. They are powerful. Impossible interpretation of the functionality is the drawback of neural network. It also faces problem in determining number of layers and it can be used in hardware as well as software. Combination of neural network and fuzzy logic known as FNN. They are universal approximators.

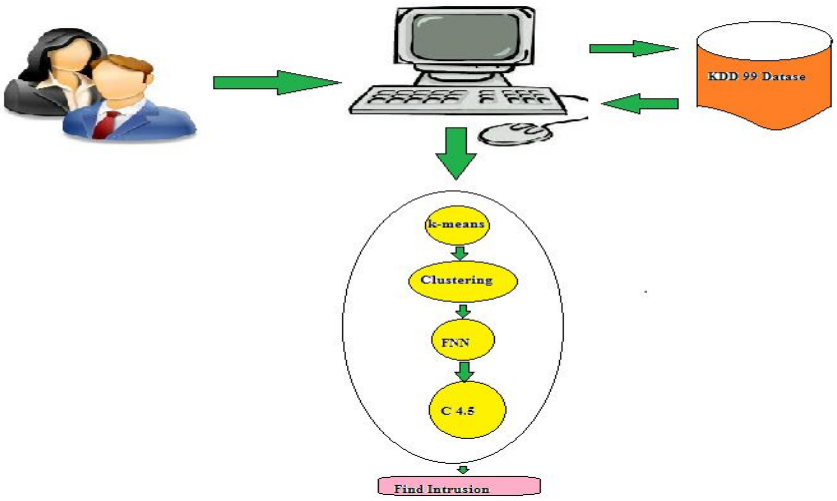

Figure 3 System Design for IDS using C 4.5

C 4.5 Algorithms: it is an algorithm used to generate a tree. It is developed by Ross Quinlan. C4.5 acts as an extension to ID3 algorithm. C4.5 generate decision tree which can be used for classification, and for this reason, C4.5 is always called as a statistical classifier.C4.5 made a number of advances to ID3. These improvements are as follow:

- They can handle both continuous and discrete attributes
- They can handle training data with missing attribute values
- They can handle attributes with differing costs.
- They Prune trees after creation

5. Results

We used 10% KDD 99 dataset for our system. We implement out system on a windows PC with i3 processor 2.30 GHz CPU and 4GB RAM. We used KDD 99 dataset for experiment. It is publically available dataset. We used evaluation metrics like sensitivity, specificity and accuracy.
Sensitivity=TP/ (TP+FN)
Specificity=TN/ (TN+FP)
Accuracy= (TN+TP)/ TN+TP+FN+FP

Table 3 Accuracy Comparison of SVM and C 4.5

Methods	Types of Attacks			
	DOS	PROBE	R2L	U2R
Using SVM	95	96	95	97
Using C4.5	97	98	97	98

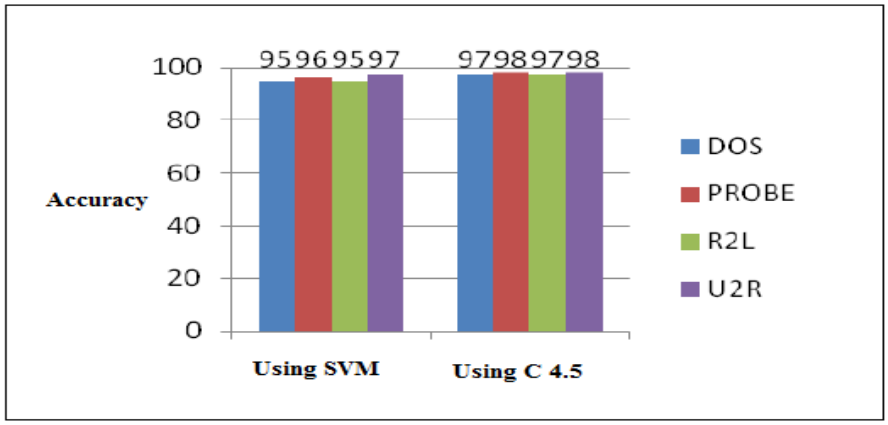

Figure 4 Graphical Representation of Comparison of Various Techniques

Proposed technique has gained a reliable peak score. We compare two systems. One is SVM and other is C 4.5.results show that C 4.5 performs better than SVM.

6. CONCLUSION

At the present time intrusion has become major concern for many organizations. It influences the security as well as privacy of the system. Hence, there is a need of very intense IDS which can detect rare attack with high accuracy detection of attack detection. In this work, we have proposed intrusion detection technique using C 4.5.it reduces the time needed to construct model in case of classification as well as increase the intrusion detection accuracy. From experiment it is prove that, when data sets are properly processed, it can overcome the drawback of extensive time required to build model.

References

[1] Miss Meghana Solanki, Mrs. Vidya Dhamdhere, "Intrusion Detection System by using K-Means clustering, C 4.5, FNN, SVM classifier", Volume 3, Issue 6, November-December 2014, ISSN 2278-6856.I.S

[2] Miss Meghana Solanki, Mrs. Vidya Dhamdhere, "Intrusion Detection Technique using Data Mining Approach: Survey", Vol. 2, Issue 11, November 2014, ISSN(Online): 2320-9801.

[3] James P. Anderson, "Computer Security Threat Monitoring and Surveillance," Technical report, James P. Anderson Co., Fort Washington, Pennsylvania. April 1980. J. Geralds, "Sega Ends Production of Dreamcast," vnunet.com, para. 2, Jan. 31, 2001. [Online]. Available: http://nl1.vnunet.com/news/1116995. [Accessed: Sept. 12, 2004]. (General Internet site)

[4] Kaining Lu Zehua Chen Zhigang Jin Jichang Guo." An Adaptive Real-Time Intrusion Detection System Using Sequences of System Call", CCECE 2003

[5] Chunhua Gu and Xueqin Zhang," A Rough Set and SVM Based Intrusion Detection Classifier", Second International Workshop on Computer Science and Engineering, 2009.

[6] Yong-Xiang Xia, Zhi-Cai Shi, Zhi-Hua Hu," An Incremental SVM for Intrusion Detection Based on Key Feature Selection" 2009 Third International Symposium on Intelligent Information Technology Application.

[7] Rung-Ching Chen, Kai-Fan Cheng and Chia-Fen Hsieh ,"Using Rough Set And Support Vector Machine For Network Intrusion Detection" International Journal of Network Security & Its Applications (IJNSA), Vol 1, No 1, 2009.

[8] Heba F. Eid, Ashraf Darwish, Aboul Ella Hassanien, and Ajith Abraham" Principle Components Analysis and Support Vector Machine" based Intrusion Detection System", IEEE 2010.

[9] Shingo Mabu, Ci Chen, Nannan Lu, Kaoru Shimada,and Kotaro Hirasawa," An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming", IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 41, No. 1, January 2011.
[10] Carol J Fung and *Jie* Zhang," Dirichlet-Based Trust Management for Effective Collaborative Intrusion Detection Networks" ,IEEE Transactions on Network And Service Management, Vol. 8, No 2, June 2011.
[11] R Rangadurai Karthick, Vipul P. Hattiwale, Balaraman Ravindran," Adaptive Network Intrusion Detection System using a Hybrid Approach ", IEEE 2012.
[12] V. Jaiganesh, "Intrusion Detection Using Kernelized Support Vector Machine With Levenbergmarquardt Learning", International Journal of Engineering Science and Technology, 2012
[13] Gholam Reza Zargar, Tania Baghaie, "Category-Based IntrusionDetection Using PCA", Journal of Information Security, 2012.
[14] Y. Chen, A. Abraham, B. Yang, "Hybrid flexible neural-tree-based intrusion detection systems", Int. J. Intell. Syst. 22 (2007) 337–352.