ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

Identifying Intrusion Detection System using Hybrid technique with Support Vector Machine

Swapna Yendole, Prof.Sujata Tuppad

M.E Student, Dept. of CSE, MSS's College of Engineering and Technology, Jalna, India

Professor, Dept. of CSE, MSS's College of Engineering and Technology, Jalna, India

ABSTRACT: This study proposed an SVM based on IDS that combines GA and SVM techanique.GA is used to pre-process the KDD Cup data set (1999) before the SVM training. The proposed system reduces training time and also allows better classification of different types of attacks. Various transaction activities such as booking airline tickets, online banking, distance learning, group discussion and so on are performed using the Internet. The realization of these activities involves the exchange of useful data that must be protected against malicious attacks. If certain malicious activities occur in the network, it is essential to alert the user to this. Detecting malicious activity is critical to protecting our data. To protect the data that is exchanged on the network, we need to implement a system that achieves faster attack detection and response accordingly. To detect malicious activity, we use the intrusion detection system (IDS) for security reasons. GA algorithm is used to provide a set of high quality, abstract featuresand reduced data to SVM for training. This system tries to increase the accuracy of the probe and u2r attacks. This system is implemented in Java.

KEYWORDS: Hybrid Method; Support Vector Machine KDD Cup 1999; Network Intrusion DetectionSystem; Support Vector Machine.

I.INTRODUCTION

As the use of the Internet is growing day to day, its security has been a focus in current research. Today, special attention has been paid to the intrusion detection system (IDS), which is closely linked to the secure use of network services. The Network Intrusion Detection System (NIDS), as an important link in network security infrastructures, aims to detect malicious activities such as denial of service attacks, port analyzes and cracking attempts in computers. Monitoring network traffic. One of the common problems with NIDS is that it only specifically detects known service or network attacks, known as misuse, using model matching approaches. On the other hand, an anomaly detection system detects attacks by first constructing normal behavior patterns and then identifies potential attacks when their behaviors are significantly deviated from normal profiles.
Much research has applied data mining techniques in the design of NIDS. One promising technique is the Support Vector Machine (SVM), whose solid mathematical foundations [Khan et al., 7] have yielded satisfactory results. SVM separates the data into several classes (at least two) by a hyperplane, and simultaneously minimizes the empirical classification error and maximizes the geometric margin. Thus, it is also known as classifiers of the maximum margin. A hierarchical grouping algorithm that is used to produce fewer significant instances from a very large set of data. With fewer significant cases, support vector machines (SVMs) can achieve a shorter training time and better classification performance.

In Misuse detection, a low false positive rate is obtained and a minor variation of known attacks can not be detected whereas Anomaly detection has a high false positive rate because it can detect new attacks. In an ideal system of intrusion detection, a high attack detection rate and a false positive rate of 0% should be present. This low rate of false positives is achieved only to the detriment of detection of minor malicious activity. Since the two elements show the complementary nature, many systems attempt to combine the two techniques where misuse detection techniques can be used as the first line of defense, whereas abnormality detection techniques can be a second .

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

Most intrusion detection systems are classified either as a network-based approach or as a host-based approach to detecting and detecting attacks. A network-based intrusion detection system performs traffic analysis on a local area network. A host-based intrusion detection system places its reference monitor in the kernel / user layer and monitors for anomalies in the system call patterns. The advantages of using network-based intrusion detection systems are no processing effect on monitored hosts, the ability to observe events at the network level and monitor an entire segment at a time. However, as complexity and network capacity increase, performance requirements for probes may become prohibitive [Chen et al., 21]. Host-based intrusion detection systems can analyze all activities on the host, including its own network activities. Unfortunately, this approach implies an impact on the performance of each monitored system [Verwoerd & Hunt, 20].
This study proposed an intrusion detection system based on SVM, hierarchical clustering and genetic algorithm. The genetic algorithm is used to eliminate the unimportant characteristics of the learning set so that the resulting SVM model can classify network traffic data more accurately. The hierarchical grouping algorithm stores fewer abstract data points in the KDD Cup 1999 dataset than the dataset. Thus, the system could significantly reduce learning time and obtain better detection performance in the resulting SVM classifier. The rest of his papers are organized like this. Section 2 presents the hierarchical grouping, the genetic algorithm and the SVM. Section 3 describes the proposed system. Section 4 shows the experimental results. Finally, Section 5 refers to the conclusion.

II.RELATED WORK

In [2] Author presents network intrusion detection using agent and SVM to improve the accuracy of detection of intrusive attacks. The network intrusion detection system consists of a data acquisition module, an intrusion detection agent and the management agent. The model used in this paper uses four SVM classifiers that classify network data into five classes: DoS, probe, U2R, R2L, and normal. The experiment is performed on the KDDCUP99 dataset to detect the precision of the attack by applying the SVM model. The same dataset is used to measure the attack detection accuracy by applying the back propagation approach. By comparing the results, one can know that the SVM detection accuracy is 0.9457 which is more compared to the back propagation which is 0.8771. From the experiment performed, the application of the technique of SVM and agent is preferable to the artificial neural network application (back propagation).
In[3] It also shows the optimal hyper plane for binary classification of data points. An optimal hyper plane should be selected such that it should maximize the distance between its nearest points that belong to the class [16]. This distance is known as margin.As SVM performs binary classification, we need a mechanism that intelligently classifies data when more than one target class is present. In order to classify data in more than one class, a multiclass classification approach is adopted. SVM performs the multi-class classification.
This paper describes the use of a raw ensemble hybrid method and SVM. Hybrid method is used in this paper where the approximate approach is used for data reduction and SVM approach is used for classification of data and detection of intrusion. Here, the output can belong to one of five classes: probe, U2R, R2L and the DoS attack data and the normal data. Before providing the data to the SVM classifier, data reduction is accomplished by applying a rough set to simplify the training data. After that, the data passes through the SVM classifier Finally, the experiment was carried out on the selected data set and the results of the experiment concluded that the detection accuracy of RS-SVM was higher (93.64%) than that of SVM (86.62%).

III.PROPOSED ALGORITHM

We focus on an automated learning model using a modified vector support machine (SVM) that combines the benefits of supervised and unsupervised learning. In addition, we propose a preliminary feature selection process using GA to select more appropriate packet fields. Now we discuss our hybrid algorithm steps which are as follows:

Step 1 - Load the kdd dataset first.

Step 2 - Preprocessing the Data
Here, process all the data in the database. The KDD CUP database "99 contains 41 functions such as dst_bytes, src_bytes, etc. Since the SVM classification uses only numerical data for testing and training, it is necessary to convert

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

the textual data into numerical values. , We assumed some numeric values for different text functionality, such as "protocol_type" functionality "tcp" as 3, "udp" as 7, and "icmp" as 9 etc as shown in the table.

Step 3- Feature Selection Algorithm (Weka SVM Decision Tree Stump)
In this work, the genetic algorithm-based approach is proposed to select the optimal characteristics of the 41 global characteristics. The selected characteristics discriminate in the predictive class when classifying for abnormality.
The steps of the algorithm are as follows:
1. Generate a random population of n chromosomes (data set of appropriate solutions for the problem)

2. Evaluate the physical condition f (x) = k (x) / sqrt (k + k (k-1) x) where k is a random number and x represents the chromosome of the population
3. Create a new population by repeating the following steps until the new population is complete,

A) Choose two relative chromosomes of a population according to their fitness (better physical condition, greater chance of being selected).
B) With a probability of crossing, the parents form a new offspring (children). If no crossing has been done, the offspring is an exact copy of the parents.
C) With a mutation probability, mutate new offspring at each locus (position in the chromosome).
D) Place new descendants in a new population.
4. Use the new generated population for a new algorithm sequence
5. If the final condition is satisfied, stop and return the best solution in the current population
6. Go to step 2.

Step 4- Selected feature
The main reason for selecting the KDD Cup 99 dataset is that it is currently the commonly used dataset that is shared by many researchers. In this dataset, 41 attributes are used in each record to characterize the behavior of network traffic. Of these 41 attributes, 38 are numeric and 3 symbolic. The characteristics present in the KDD dataset are grouped into three categories and are discussed below.

A. Basic Functions: Basic features include all attributes extracted from a TCP / IP connection. These functions are extracted from the packet header and include bytes src, dst_ bytes, protocol etc.

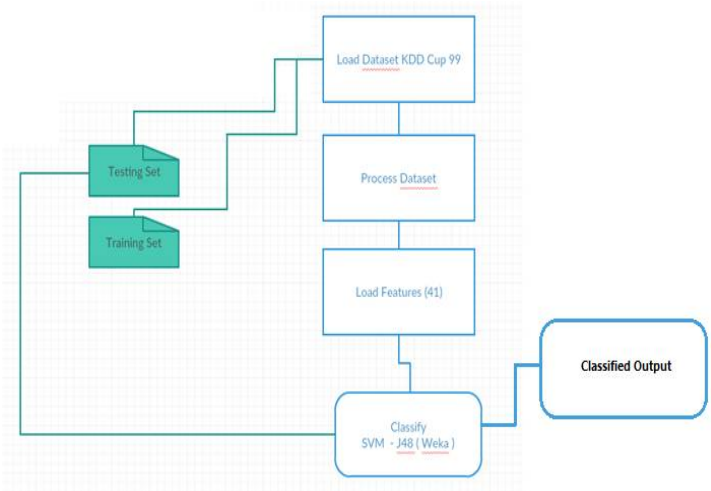

Figure 1 Architectural Diagram

B. Content features: These functions are used to evaluate the payload of the original TCP packet and to look for suspicious behaviors in the payload part. This includes features such as the number of failed logon attempts, the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

number of file creation operations, and so on. In addition, most R2L and U2R attacks do not have frequent sequential sequences. This is because DoS and Probing attacks involve many connections to certain hosts in very short time, but R2L and U2R attacks are embedded in the data portions of the packets and usually involve a single connection. To detect these types of attacks, content-based features are used.

C. Traffic functions: These functions include functions calculated according to a window interval and are divided into two categories
I) "Same host" functions: These functions are derived only by examining the connections of the last 2 seconds which have the same destination host as the current connection and calculating the statistics relating to the behavior of the protocol, the service and so on.
B. Content features: These functions are used to evaluate the payload of the original TCP packet and to look for suspicious behaviors in the payload part. This includes features such as the number of failed logon attempts, the number of file creation operations, and so on. In addition, most R2L and U2R attacks do not have frequent sequential sequences. This is because DoS and Probing attacks involve many connections to certain hosts in very short time, but R2L and U2R attacks are embedded in the data portions of the packets and usually involve a single connection. To detect these types of attacks, content-based features are used.

C. Traffic functions: These functions include functions calculated according to a window interval and are divided into two categories
II) "Same host" functions: These functions are derived only by examining the connections of the last 2 seconds which have the same destination host as the current connection and calculating the statistics relating to the behavior of the protocol, the service and so on.

Step 5- Classification algorithm

We have divided the user's behavior into two classes, namely attack and normal, where the user's behavior is the collection of different attacks belonging to the five classes such as
1 Normal-- Normal
2 DoS - apache2, back, earth, mailbomb, neptune, pod, processtable, smurf, tear, udpstrom

3 Probe - ipsweep, mscan, nmap, portsweep, saint, satan
4 R2L - ftp_write, guess_passwd, imap, multihop, named, phf, sendmail, spy, snmpgetattack, snmpguess, warezclient, warezmaster, worm, xlock, xsnoop
5 U2R - buffer_overflow, httptunnel, loadmodule, perl, ps, rootkit, sqlattack, xtern

The purpose of our SVM experiment is to differentiate the normal behavior and attack behavior of the user. In our experiments, normal data is classified as -1 and all attacks are categorized as +1.

The design of the basic input data and the output data areas are given as follows:
(Xi, yi), ..., (xn, yn), x Rm, y {+ 1, -1}
Where xi, yi, ..., (xn, yn) are train data, n is the number of samples, m is the input vector and y is inserted into the category +1 or -1 respectively. On the linear problem, a hyperplane can be divided into two categories. The formula of the hyperplane is:
$(W.x) + b = 0$
The categories are:
$(W.x) + b \geq$ if yi = +1
$(W.x) + b \leq$ if yi = -1

Step-6 classification result

Step 7- Anomaly detect.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

IV.PSEUDO CODE

Step 1: Load the kdd dataset first
Step 2: Preprocessing the Data.
Step 3: Feature Selection Algorithm (Weka SVM Decision Tree Stump)
Step 4: Selected feature
Step 5: Classification algorithm
Step 6: classification result
Step 7: Anomaly detect.
Step 8: End

V.RESULTS

The KDD CUP 1999 dataset is an extension of the DARPA 98 dataset with a set of additional functionalities. However, it does not contain some basic information about network connections (for example, start time, IP addresses, ports, etc.). The dataset was mainly built for the purpose of applying data mining algorithms. Therefore, we used this dataset as a test bench for our algorithms. The dataset contains approximately 4900000 simulated intrusion records. The simulated attacks fell into one of four categories: DOS, R2L, U2R and PROBE. There are a total of 22 attack types and 41 attributes (34 continuous and seven categorical). It appears that the data set is too large. However, generally only 10% of the subset is used to evaluate the performance of the algorithm. The 10% subset contains the 22 types of attack. It consists of all low frequency attack recordings and 10% of normal recordings and high frequency attack recordings, such as smurf, neptune, portsweep and satan. These four types of attack records occupy 99.51% of all KDDCUP99 data and 98.45% of the 10% subset [Sulaimana&Muhsinb, 11].
In the 1999 KDD Cup dataset, there are 4,898,431 and 311,029 records in the training and test data sets. In this dataset, attack records were classified into four categories: DoS, U2R, R2L and Probe. In the 19.85% training set were normal traffic and rest was attack traffic while for the test data set it contained 19.48% as normal traffic and rest were attack traffic . There are 41 quantitative and qualitative characteristics in each record of the KDD dataset [Patcha& Park, 12].
The Results of our proposed system are as follows:

Table1 Results

Type of traffic	TPR	TNR	FPR	FNR	Accuracy	Precision
Dos	0.9867	0.7127	0.2873	0.0133	0.8160	0.6751
Normal	0.6000	0.9840	0.0160	0.4000	0.9130	0.8952
Probe	0.8162	0.9951	0.0049	0.1838	0.9620	0.9742
R2L	0.5940	0.9739	0.0261	0.4060	0.8850	0.8742
U2R	0.5789	1.000	0.000	0.4211	0.9920	1.000

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Table 2: Number of Attack Instances

Types of Attack	Attack Instances
Dos	377
Normal	185
Probe	185
R2L	234
U2R	19

From the result, we have seen that fp of u2r and the probe is less due to what the precision of this attack is more, and also in about 1000 cases, about 18.5% and 1.9% are attacks Of probe and u2r respectively.

VI. Conclusion and Future Work

In this study, an SVM-based intrusion detection system that combines genetic algorithm j48 and SVM technique. The J48 algorithm is used are used for the pre-processing of the data and feature selection. Both j48 and SVM techniques provides highly qualified, abstract and reduced data to SVM training.. The well-known KDD CUP 1999 dataset was used to evaluate the proposed system. Compared to other intrusion detection systems, this system has shown better performance in detecting various attacks. Future work is to apply SVM with other data pre-processing techniques.

References

[1] D. Anderson, T.F. Lunt, H. Javitz, A. Tamaru & A. Valdes (1995), "Detecting Unusual Program Behavior using the Statistical Component of the Next-generation Intrusion Detection Expert System (NIDES)", Menlo Park, CA, USA: *Computer Science Laboratory, SRI International*. SRI-CSL-95-06.

[2] A. Satsiou, M. Doumpos& C. Zopounidis, "Genetic Algorithms for the Optimization of Support Vector Machines in Credit Risk Rating".

[3] K. Burbeck& N.Y. Simmin (2007), "Adaptive Real-Time Anomaly Detection with Incremental Clustering", *InformationSecurity Technical Report*, Vol. 12, No. 1, Pp. 56–67.

[4] W. Chimphlee, A.H. Abdullah, M.N. Md Sap, S. Srinoy& S. Chimphlee (2006), "Anomaly-based Intrusion Detection using Fuzzy Rough Clustering", *Proceedings of the InternationalConference on Hybrid Information Technology (ICHIT'06)*.

[5] T.S. Chou, K.K. Yen & J. Luo (2008), "Network Intrusion Detection Design using Feature Selection of Soft Computing Paradigms", *International Journal of ComputationalIntelligence*, Vol. 4, No. 3, Pp. 196–208.

[6] R. Fei, L. Hu & H. Liang (2008), "Using Density-based Incremental Clustering for Anomaly Detection", *Proceedingsof the 2008 International Conference on Computer Science and Software Engineering*, Vol. 3, Pp. 986–989.

[7] L. Khan, M. Awad& B. Thuraisingham (2007), "A New Intrusion Detection System using Support Vector Machines and Hierarchical Clustering", *The International Journal on VeryLarge Data Bases*, Vol. 16, No. 4, Pp. 507–521

[9] J.H. Lee, S.G. Sohn, B.H. Chang & T.M. Chung (2009), "PKG-VUL: Security Vulnerability Evaluation and Patch Framework for Package-based Systems", *ETRI Journal*, Vol. 31, No. 5, Pp. 554–564.K.Y. Lam, L. Hui& S.L. Chung (1996),"A Data ReductionMethod for Intrusion Detection", *Systems Software*, Vol. 33,Pp. 101–108.

[10] G. Helmer&G. Liepins (1993),"Statistical Foundations ofAudit Trail Analysis for the Detection of Computer Misuse", *IEEE Transactions on Software Engineering*, Vol. 19, Pp. 866–901.

[11] N. Sulaimana& O.A. Muhsinb (2011), "A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment", *Procedia Computer Science*, Vol. 3, Pp. 1237– 1242.

[12] A. Patcha& J.M. Park (2007), "Network Anomaly Detection with Incomplete Audit Data", *Computer Networks*, Vol. 51, No. 13, Pp. 3935–3955.

[13] S. Horng, M. Su, Y. Chen, T. Kao, R. Chen, J. Lai & C.D. Perkasa (2011), "A Novel Intrusion Detection System based on Hierarchical Clustering and Support Vector Machines", Expert Systems with Applications, Vol. 38, No. 1, Pp. 306–313.

[14] P. Soto (2001), "The New Economy Needs New Security Solutions", http://www.xcf.berkeley.edu/~paolo/ids.html.

[15] T. Pietraszek& A. Tanner (2005), "Data Mining and Machine Learning towards Reducing False Positives in Intrusion Detection", Information Security Technical Report, Vol. 10, No. 3, Pp. 169–183.

[16] A.N. Toosi& M. Kahani (2007), "A New Approach to Intrusion Detection based on an Evolutionary Soft Computing Model using Neuro-Fuzzy Classifiers", Computer Communications, Vol. 30, Pp. 2201–2212

[17] T. Verwoerd & R. Hunt (2002), "Intrusion Detection Techniques and Approaches", Computer Communications, Vol. 25, Pp. 1356–1365.

[18] W.-H. Chen, S.-H. Hsu & H.-P. Shen (2005), "Application of SVM and ANN for Intrusion Detection", *Computers &Operations Research*, Vol. 32, No. 10, Pp. 2617–2634.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

[19] T. Zhang, R. Ramakrishnan& M. Livny (1996), "BIRCH: An Efficient Data Clustering Method for Very Large Databases", *Proceedings of the ACM SIGMOD (SIGMOD'96)*, Pp. 103–114.

BIOGRAPHY

SwapnaR.Yendole is a Student of Master of Engineering in the Computer Science and Engineering Department, Matsyodari College of Engineering & Technology, Babasaheb Ambedkar Marathwada University, She Received Bachelor of Engineering(BE) in 2013 from BAMU ,Aurangabad, Maharashtra, India