Implementing Intrusion Detection System Using Feature Selection and Classification Technique

[1]P.V.N Rajeswari, [2]Nataraja Sirisha

[1]Assoc. Professor in Dept. of CSE, Visvodaya Engineering College, Kavali, and Andhra Pradesh, India.

[2]M.Tech, Dept. of CSE, Visvodaya Engineering College, Kavali, and Andhra Pradesh, India.

Abstract – *With the growth of Internet, there has been a tremendous increase in the number of attacks and therefore Intrusion Detection Systems (IDS's) has become a main stream of information security. The purpose of IDS is to help the computer systems to deal with attacks. This anomaly detection system creates a database of normal behaviour and deviations from the normal behaviour to trigger during the occurrence of intrusions. Based on the source of data, IDS is classified into Host based IDS and Network based IDS. In network based IDS, the individual packets flowing through the network are analyzed where as in host based IDS the activities on the single computer or host are analyzed. The feature selection used in IDS helps to reduce the classification time. In this paper, the IDS for detecting the attacks effectively has been proposed and implemented. For this purpose, a new feature selection algorithm called Optimal Feature Selection algorithm based on Information Gain Ratio has been proposed and implemented. This feature selection algorithm selects optimal number of features from KDD Cup dataset. In addition, two classification techniques namely Support Vector Machine and Rule Based Classification have been used for effective classification of the data set. This system is very efficient in detecting DoS attacks and effectively reduces the false alarm rate. The proposed feature selection and classification algorithms enhance the performance of the IDS in detecting the attacks.*

***Keywords* :** Intrusion Detection; Information Gain; Support Vector Machine; Feature Selection Technique and Classification

I. INTRODUCTION

Computers have been networked together with very large user source and so security has been a vital concern in many areas. With the rapid growth of internet communication and availability of tools to intrude the network, security for network has become indispensable. Current security policies do not sufficiently guard the data stored in the databases. Many other technologies like firewalls, encryption and authorization mechanisms can offer security, but they are still sensitive for attacks from hackers who takes advantage of the system flaws [3] .To protect these systems from being attacked by intruders, a new Intrusion Detection System has been proposed and implemented in this project work, which combines a simple feature selection algorithm and SVM technique to detect attacks in [9].

Using KDD cup data set and Data Mining extract the hidden predictive information from large Databases. It is a powerful new technology with great potential that helps companies focus on the most important information in their data warehouses. Data mining can be applied to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Recently, internet has become a part of daily life. The current internets based on information processing systems are prone to different kind of threats which lead to various types of damages resulting in significant losses. Therefore, the importance of information security is evolving quickly.

The most basic goal of network security is to develop defensive networking systems which are secure from unauthorized access, using disclosure, disruption, modification, or destruction in [10]. Moreover, network security minimizes the risks related to the main security goals like confidentiality, integrity and availability.

II. BACKGROUND WORK

In recent times, network security has been the subject of many research works with advent of internet. There are many works in the literature that discuss about Intrusion Detection System. IDSs are used to detect the attacks made by intruders [2]. Sindhu et al proposed a genetic based feature selection algorithm for minimizing the computational complexity of the classifier. Lee et al proposed an adaptive data mining approach for intrusion detection in which association rules and frequent episodes derived from audit data are used as the basis for the feature selection process. Xiang and Lim proposed a misuse IDS using multiple level hybrid classifier. Moradi and Zulkernine [4] presented IDS that uses ANN for effective intrusion detection.

One of the limitations of their approach is that it increases the training time. Sarasamma et al proposed a novel multilevel hierarchical Kohonen networks to detect intrusions in networks. In their work, they randomly selected data points from KDD Cup 99 to train and test the classifier. Jianping Li et al [6] proposed a new method based on Continuous Random Function for selecting appropriate feature sets to perform network intrusion detection. There are many classification algorithms based on SVM that are found in the literature for IDS. For example, an algorithm called Tree Structured Multiclass SVM had been proposed by Snehal A. Mulay et al for classifying data effectively. There are many works in the literature that discuss about preprocessing. Most of the real life problems definitely need an optimal and acceptable solution rather than calculating them precisely at the cost of degraded performance, time and space. The feature selection search started with null set where features were added one by one or it was started with a full set of features where features were eliminated one by one. Li et al proposed a wrapper based feature selection algorithm in order to develop an IDS. The feature selection algorithm proposed by Geetha Raman ideals with the statistical method for analyzing the voluminous KDD Cup dataset.

There are many works in the literature that discuss about classification techniques and tools. Support Vector Machines (SVM) were the classifiers which were originally designed for binary classification. Debar et al developed a Neural Network model for IDS. Du Hongle et al proposed an improved v-FSVM through introduction membership to each data point. Dewan Md. Farid proposed a new learning approach for network intrusion detection using naïve Bayesian classifier and ID3 algorithm is presented, which identifies effective attributes from the training dataset, calculates the conditional probabilities for the best attribute values, and then correctly classifies all the examples of training and testing dataset.

The SVM based intrusion detection system combines a hierarchical clustering algorithm, a simple feature selection procedure, and the SVM technique.

Data Preparation Subsystem

Data Collector

The Data collection agent collects the records from the KDD’99 cup data set. This data is sent to the data preprocessing module for pre-processing the data. The records collected from the KDD cup dataset may be a normal data or an attacked data.

Pre-processing Module

Pre-processing techniques are necessary for data reduction since it is quiet complex to process huge amount of network traffic data with all features to detect intruders in real time and to provide prevention methods.

Classification Subsystem

Rule Based Classifier

In this system, decisions on anomaly intrusion detection and prevention are improved by the application of rules fired using the rule system invoked by the intelligent agents. The main advantage of using rules with knowledge base is that it helps to perform effective decision making on intrusions.

Support Vector Machine

SVM is the learning machine that can perform binary classification and regression estimation tasks. They are becoming increasingly popular as a new paradigm of classification and learning because of two important factors. First, unlike the other classification techniques, SVM minimizes the expected error rather than minimizing the classification error. Second, SVM employs the duality theory of mathematical programming to get a dual problem that admits efficient computational methods.

III. PROPOSED WORK

Proposed Algorithm for Optimal Feature Selection

This algorithm has been developed by calculating Information Gain Ratio for attribute selection. In order to achieve this, the data set D is divided into n number of classes Ci. The attributes Fi having maximum number of non-zero values are chosen by the agent and the Information Gain Ratio (IGR) is computed using equations:

$$info(D) = -\sum_{j-1}^{m}\left[\frac{freq(C_j,D)}{|D|}\right]\log_2\left[\frac{freq(C_j,D)}{|D|}\right] \quad (1)$$

$$info\ (F) = \sum_{i=1}^{n}\left[\frac{|F_i|}{|F|}\right] * info(F_i) \quad (2)$$

$$IGR\ (Ai) = \left[\frac{info(D)-info(F)}{info(D)+info(F)}\right] * 100 \quad (3)$$

The steps of the optimal feature selection algorithm are as follows.

Algorithm: Intelligent Agent based Attribute Selection Algorithm

Input: Set of 41 features from KDD'99 Cup data set

Output: Reduced set of features R

1. Select the attributes which have variation in their values.
2. Calculate the Info (D) values for the selected attributes using the equation 1.
3. Select the attributes which have maximum
4. number of non-zero values.
5. Calculate the Info(F) value for the attributes selected in step 3 using the equation 2.
6. Calculate the IGR value using the equation 3.
7. Depending on the IGR value, select the attributes.

The OFS algorithm has selected 10 important features for effectively detecting the attacks and to reduce the computation time.

The pseudo code for optimal feature selection is given below.

Input the data set d

```
for each column in the data set
Select non-varying columns
end for
for each non-varying columns
calculate frequency of each value in the data set
calculate info(d)
end for
for each column with maximum no. of non-zero values
calculate frequency of each value
calculate info(f)
end for
for each column
calculate IGR value
end for
```

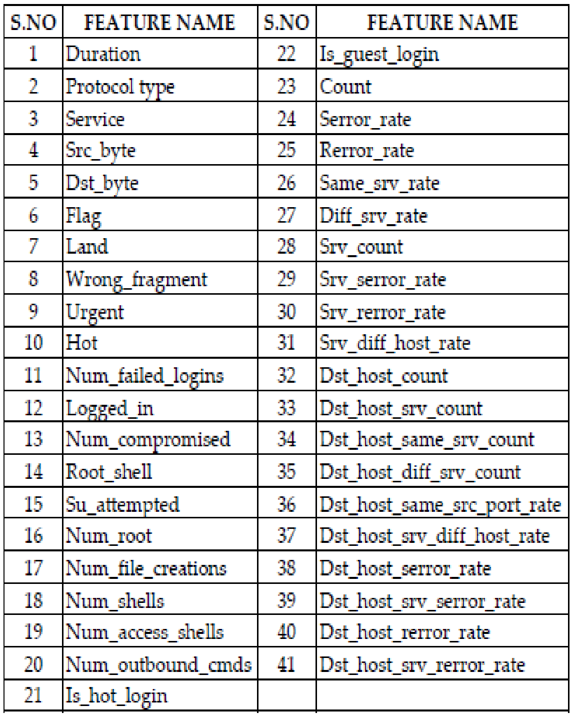
In this chapter, the algorithm to implement the OFS is presented. The next chapter discusses the implementation and analysis the performance results of the project.

IV. IMPLEMENTATION

a) Optimal Feature Selection

The normal feature selection algorithms take large computation time for calculating IGR values. Hence in this work, a new feature selection algorithm called Optimal Feature Selection algorithm that reduces the time taken for computation is proposed and implemented. This algorithm calculates the Information Gain Ratio (IGR) value for the varying attributes in the data set. It performs column reduction based on the IGR value. OFS increases the accuracy in detection and reduces the false alarm rates. The simulated attacks fall in one of the following four categories namely, Denial of Service (DoS), User to Root (U2R), Remote to Local (R2L) and Probe attack.

Table 1 : 41 features in kdd'99 dataset



S.NO	FEATURE NAME	S.NO	FEATURE NAME
1	Duration	22	Is_guest_login
2	Protocol type	23	Count
3	Service	24	Serror_rate
4	Src_byte	25	Rerror_rate
5	Dst_byte	26	Same_srv_rate
6	Flag	27	Diff_srv_rate
7	Land	28	Srv_count
8	Wrong_fragment	29	Srv_serror_rate
9	Urgent	30	Srv_rerror_rate
10	Hot	31	Srv_diff_host_rate
11	Num_failed_logins	32	Dst_host_count
12	Logged_in	33	Dst_host_srv_count
13	Num_compromised	34	Dst_host_same_srv_count
14	Root_shell	35	Dst_host_diff_srv_count
15	Su_attempted	36	Dst_host_same_src_port_rate
16	Num_root	37	Dst_host_srv_diff_host_rate
17	Num_file_creations	38	Dst_host_serror_rate
18	Num_shells	39	Dst_host_srv_serror_rate
19	Num_access_shells	40	Dst_host_rerror_rate
20	Num_outbound_cmds	41	Dst_host_srv_rerror_rate
21	Is_hot_login		

Calculation of info (D)

The information gain criterion is derived from information theory. The essential idea of information theory is that the information conveyed by a message depends on the probability and can be measured in bits as minus the logarithm of base 2 of that probability. Suppose we have a dataset D with q classes C1,…Cn. Suppose further that we have a possible test x with m utcomes that partitions D into m subsets D1,…,Dm. For a numeric attribute, m=2, since we only perform binary split. The probability that is selected one record from the set D of data records and announce that if belongs to some class Cj is given by ,

$$\sum_{j=1}^{m}\left[\frac{freq(C_j,D)}{|D|}\right] \qquad (4)$$

Where freq (Cj, D) represents the number of data records(points) of the class Cj in D, while |D| is the total number of data records in D. So the information that is convey is

$$-\log_2\left[\frac{freq(C_j,D)}{|D|}\right] bits \qquad (5)$$

To find the expected information needed to identify the class of a data record in D before partitioning occurs, summation is performed over the classes in proportion to their frequencies in D, in eq. (1). Now, suppose that the dataset D has been partitioned in accordance with the m outcomes of the test x. The expected amount of information needed to identify the class of a data record in D after the partitioning has occurred, can be found as the weighted sum over the subsets, as shown in eq. (2). The information gained due to the partition is:

$$Gain(Ai) = info(D) - info(F) \qquad (6)$$

Clearly, it is necessary to maximize the gain. The gain criterion is to elect the test or cut the maximizes the gain to partition the current data.

V. CONCLUSION

In this work, a new IDS has been proposed and implemented by combining an Optimal Feature Selection (OFS) algorithm and two classification techniques for securing the system. The computation time taken for detecting and classifying the records using all the forty one features of the KDD'99 cup data set is observed to be large. The proposed feature selection algorithm selects only the important features that help in reducing the time taken for detecting and classifying the records. Further the rule based classifier and SVM help achieve a greater accuracy. The main advantage of the proposed IDS is that it reduces the false positive rates and also reduces the computation time.

REFERENCES

[1] Daramola O. Abosede, Adetunmbi A. Olusola, AdeolaS. Oladele,. "Analysis of KDD'99 Intrusion Detection Dataset for Selection of Relevance Features", Proceedings of the World Congress on Engineering and Computer Science, Vol. I, October 20-22, 2010.

[2] Devale.P.R,Garje.G.V.,SnehalA.Mulay, 2012. "Intrusion Detection System using Support Vector Machine and Decision Tree ", International Journal of Computer (0975 – 8887), Vol. 3, June 2010.

[3] Debar, H., Becker, M. and Siboni, D. "A Neural Network Component for an Intrusion Detection System",IEEE Symposium on Research in Computer Security and Privacy, pp. 240-250, 1992.

[4] Du Hongle, Teng Shaohua and Zhu Qingfang, "Intrusion detection Based on Fuzzy support vector machines", International Conference on Networks Security, Wireless Communications and Trusted Computing, pp. 639-642, 2009.

[5] Dewan Md. Farid, Jerome Darmont, Nouria Harbi, Nauyen HuuHoa, Mohammad Zahidur Rahman. "Adaptive Network Intrusion Detection Learning: Attribute Selection and Classification", International Conference on Computer Systems Engineering, version 1 - 19, 2010.

[6] Farid D.M, Jerome Dormont, NouriaHarbi, Nguyen HuuHoa and Rahman, M.Z. "Adaptive Network Intrusion Detection Learning: Attribute Selection and Classification", International Conference on Computer Systems Engineering, Version 1, pp. 321-337, 2010.

[7] Geetha Ramani R, Siva Sathya S, Sivaselvi K. "Discriminant Analysis based Feature Selection in KDD Intrusion Dataset", International Journal of Computer Applications (0975 – 8887),Vol. 31, No.11, 2011.

[8] Leng J, Valli C, and Armstrong L. "A Wrapper-based Feature Selection for Analysis Large Data Set", Proceedings of 2010 3rd International Conference onand Electrical Engineering (ICCEE), pp. 167-170, 2010.

[9] Moradi M and Zulkernine M "A Neural Network based System for Intrusion Detection and Classification of Attacks", Proceedings of IEEE International Conference on Advances in Intelligent Systems – Theory and Applications, Luxembourg, Vol. 148, pp. 1-6, 2004.

[10] Sarasamma S., Zhu, Q. and Huff, J. "Hierarchical Kohonen Net for Anomaly Detection in Network Security", IEEE Transactions on System, Man, Cybernetics, Part B, Cybernetics, Vol. 35, No. 2, pp. 302-312, 2005.

[11] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE

Transactions on Neural Networks 5 (4) (1994) 537–550.

[12] F. Amiri, M. Rezaei Yousefi, C. Lucas, A. Shakery, N. Yazdani, Mutual information-based feature selection for intrusion detection systems, Journal of Network and Computer Applications 34 (4) (2011) 1184–1199.

[13] A. Abraham, R. Jain, J. Thomas, S. Y. Han, D-scids: Distributed soft computing intrusion detection system, Journal of Network and Computer Applications 30 (1) (2007) 81–98.

[14] S. Mukkamala, A. H. Sung, Significant feature selection using computational intelligent techniques for intrusion detection, in: Advanced Methods for Knowledge Discovery from Complex Data, Springer, 2005, pp. 285–306.

[15] S. Chebrolu, A. Abraham, J. P. Thomas, Feature deduction and ensemble design of intrusion detection systems, Computers & Security 24 (4) (2005) 295–307.

AUTHORS

P.V.N Rajeswari has received her B.Tech in Computer Science & Engineering and M.Tech degree in Computer science & Engineering from Andhra University in 2004 and Allahabad University in 2006 respectively. She is also pursuing PhD in Andhra University. She is dedicated to teaching field from the last 13 years. She has guided 18 P.G and 42.UG students. Her research areas included Data Mining., Machine Learning and Big Data. At present she is working as Associate Professor in Visvodaya Engineering College, Kavali, Andhra Pradesh, India.

Nataraja Sirisha has received her B.Tech in Information Technology from Brahmas Institute of Engineering College, Nellore affiliated to JNTU, Anantapur in 2013 and pursuing M.Tech degree in Computer Science and Engineering in Visvoday Engineering College, Kavali, A.P affiliated to JNTU, Anatapur in (2015-2017).