ISSN: 0975-8585

esearch Journal of Pharmaceutical, Biological and Chemical Sciences

Compression Based Classification Technique for Efficient Intrusion Detection.

D Nivedha, and G Manikandan*.

School of Computing, SASTRA University, Thanjavur, Tamilnadu, India

ABSTRACT

Intrusion detection system is application software it monitors the network for any abnormal or malicious activities.To minimize time taken and cost spent for detection we need to minimize total volume of a data model that was used in the process of intrusion detection. KDD99 Dataset is taken for experimental purpose. In the existing systems, PCA algorithm is used for selecting attributes. K-means clustering is employed to cluster the data as intruders and normal users. Only a particular cluster is selected for finding intrusion and so, there is data loss during clustering. OneRule algorithm is simple and easy method for feature selection.The Fuzzy class association rule mining is applied to Source bytes and destination bytes through which Rule extraction is done to find the normal users and intruders.C4.5 decision tree classification algorithm is used to construct decision tree and provides 98.4% accuracy.

Keywords: Intrusion Detection, Feature Selection, Classification, One Rule

Corresponding author

ISSN: 0975-8585

Introduction

Intrusion is a malicious activity provided by the unauthorized user to attack the authorized user information in a system or a network. The intruders try to steal the information using various types of attacks like DOS attack, probing, remote to local attack, user to root attacks. Network intrusion detection system is located in the point or points inside the network for monitoring traffic which is received from one device to another device and monitor the traffic which is sent from one device to another through a network.NIDS main functionality is monitoring traffics, it scans firewalls, scans network server, scan live traffics and it does not replaces the firewalls. It analyzes the traffic passing on subnet, and matches traffic which is passed in subnet to library of known attacks .Host Intrusion Detection System runs in a single host in a network [1-2]. HIDS which monitors the incoming and outgoing packets from any device, and alerts the user or admin management if an malicious activity is detected. In HIDS, applications like firewalls spy-ware detection program which are anti-thread are installed on each and every network. At server anti-thread software is installed. HIDS detects the potential breach, logs the data and send alerts signals to management. The connections are terminated is said as IPS.It makes a record of existing system files and compares it with the previous record.Once the malicious activity is identified or abnormal activity is sensed, alarm or any alert signal is sent to the management organization [3-4].

Existing System

The most effective and direct conditions to select features for detecting intrusion is Principle component analysis, but lacks in intelligence [5]. Some methods were proposed for feature selection such as graphic visualization, fuzzy C means, gradually removal method. Although this leads to reduction in time cost for detection, but detection accuracy is not assured. Some data sampling method has the drawback of losing the useful data.To identify abnormal activity in network communication data at early stage anomaly detection has potential power to detect unknown attacks than signature-based detection[6].Anomaly detection includes constructing a structure on training data followed by creating a data model for detection.To minimize time taken and cost spent for detection we need to minimize total volume of a data model that was used in the process of intrusion detection.K-means clustering is employed for clustering the dataset into 'n' cluster.Only a particular set of clusters alone taken for attributes,there may be loss of data during clusters. Classification technique predicts the target class accurately for each set of attributes.ID3 decision tree deals with only binary attributes [7-8].Execution time is higher gradually as the data size increases. Optimal solution is not assured for decision tree generated from ID3.

PROPOSED SYSTEM

The intrusion in network communication data is to be identified with less duration, the attributes are reduced and useful attributes alone is taken for intrusion detection. One Rule Algorithm is applied to select useful attributes from the dataset. Here, the dataset is composed of 42 attributes and frequency table is constructed for each attribute with respect to class labels in the dataset. The attributes with the minimum frequency rate is alone selected and considered as useful attributes (in this case, 20 attributes). After the intrusion detections, using class association rule mining the data is classified for better representation. Decision tree algorithm C4.5 is applied for classification of normal and intruder with respect to attribute.

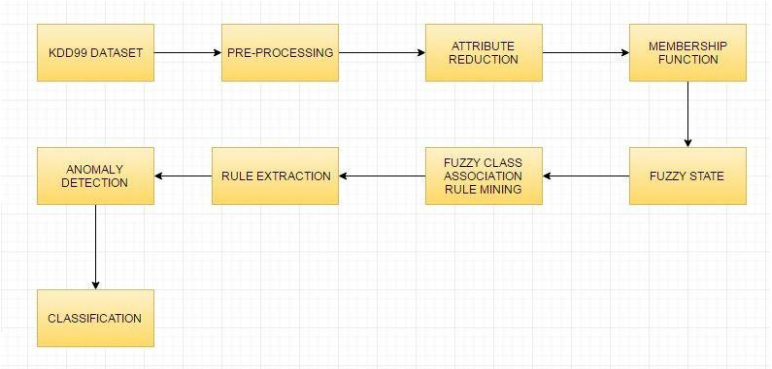

Figure 1: Architecture for Proposed System

ISSN: 0975-8585

The intruders and normal users are classified and sub divided using C4.5 algorithm. The intruders are separated based on the protocol id range. Normal users and intruders are subdivided in tree format based on protocol id range. The C4.5 provides a clear classification which predicts the output from the given input attribute with respect to corresponding class label. The class labels are intruder and normal users. The C4.5 algorithm provides the decision tree, splits the set of attribute to subsets. The Normalized Value with highest information gain is selected to form a decision. Finally intruders and normal users are classified efficiently.

METHODOLOGY AND APPROACH

Each framework has its own methodologies and approaches that are carried out by them self. The system follows some steps in order get the best output and they are:

Step 1: Preprocessing

The data is extracted from KDD99 data set. The data contains list of files called 'list files', which helps in finding a connection in a network such as Service type, source IP address, Destination IP address, Protocol type, duration , Source, destination port and the type of attacks. The preprocessor preprocess the input data and produces output which is taken as input for feature reduction. Preprocessing eliminates the missing values in the attributes.

Step 2: Attribute Selection

The frequency table is constructed for each attribute and then corresponding error rate is calculated. The attributes with minimum error rate are considered as useful attributes. It is found using One Rule Algorithm. The total number of attribute is reduced efficiently.

Step 3: Class Association Rule

The association rules are generated by given attribute conditions with respect to class labels.The step by step procedure for observing tuples is as follows.

- The database contains numerous tuples, the tuple in first position is selected it is read and a node transitions takes place from P1 processing node.
- Then, if Yes-named branch is picked up, the present node is transformed to next judging node.
- If No-named branch is chosen, P2 is transformed to processing node for calculating other rules.
- Repeat the procedure until the node transformation start from, processing node P, which is at last were finished.

After completely examining first row at the database, the second row is taken for processing and the node transformation is picked up from the P1 processing node once again. Finally, all the rows are observed by repeated process of node transformation. The number of features *(A1, A2, A3….)* in the database which equals number of judgement function *(J1, J2, J3…).*

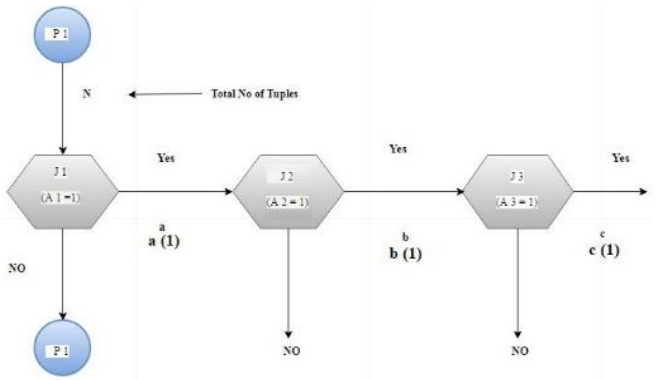

Figure 2: Node Transformation to identify the class-association rule

ISSN: 0975-8585

Step 4: Sub attribute Utilization

We introduced Sub attribute utilization method to hold continuous, binary, discrete feature types of attributes for data completeness. With respect to the judgement function, binary attributes are subdivided in to sub attributes. The symbolic feature is partitioned into two sub attributes, while the continuous feature is also subdivided into three sub attributes containing the values which are represented through a linguistic terms (lower, middle, higher) of predefined fuzzy membership function for every continuous features. In the proposed methodology, all values such as from 0 and 1 for binary features and text values are considered for symbolic features.

Step 5: Extracting Rule by Genetic Network Programming with Fuzzy Membership Function

GNP observes the features of tuples at the judgement nodes and calculates the association rules measurements only at the processing node. The operation of judgement node is to judge the value of sub attributes which are assigned, e. g., Land = 0, Protocol = UDP etc. In a single rule continuous attribute values and discrete attribute values are successfully combined using sub attribute utilization with fuzzy class-association rule based on GNP. The beginning point of class association rules is processing node P1. The Yes-named branch of the judgement node is linked with other next judgement node, while the No-named branch is linked with corresponding next processing node.

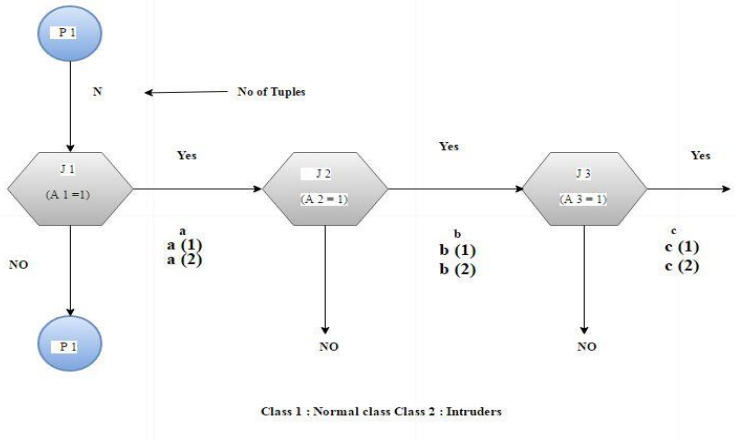

Figure 3: Node Transformation in Fuzzy Association Rule Mining

The number of rows shifting to Yes-named branch is denoted by a, b, and c. The training data contains both normal connections and anomaly connections.

EXPERIMENTAL RESULT AND ANALYSIS

The implementation of algorithm output is shown below. For experimental purpose we used KDD99 dataset which is an intrusion detection contest data. First the pre- processed data is trained with algorithm. Dataset is partitioned into two parts. The compressed data model is generated in the first part and with remaining data, compressed model is tested. One Rule algorithm is employed for selecting the useful attributes. Class association rule mining techniques is applied to identify normal user and intruders in the network communication data. Classification is the technique used for predicting the target class label for each case in the data. C4.5 classification algorithm is used to construct decision tree.Information gain is employed for splitting criteria. For making decision, attribute with highest information gain is selected. C4.5 classification for intrusion detection provides the differentiation between Intruders and normal user. Using class label the differentiation is done. The protocol duration id value with respect to class label is given as input for decision tree. The predicted output gives the clear representation of intrusions. Detection of attack is measured using metrics such as False positive (FP), False negative (FN), True Positive (TP), True Negative (TN). The efficiency of intrusion detection system is calculated using detection rate, false alarm rate, accuracy and it is represented in the following tables.

ISSN: 0975-8585

Parameters	C4.5	ID3
Accuracy	98.4 %	93 %
False Alarm Rate	9.5 %	6.12 %
Detection Rate	99.37	86

Table 1: RESULTS OF C4.5, ID3 ALGORITHMS

Parameters	C4.5	ID3
Accuracy	96 %	92.45 %
False Alarm Rate	9.80 %	5.2 %
Detection Rate	100	91.3

Table 2: RESULTS OF C4.5, ID3 ALGORITHMS FOR 20% OF RECORDS

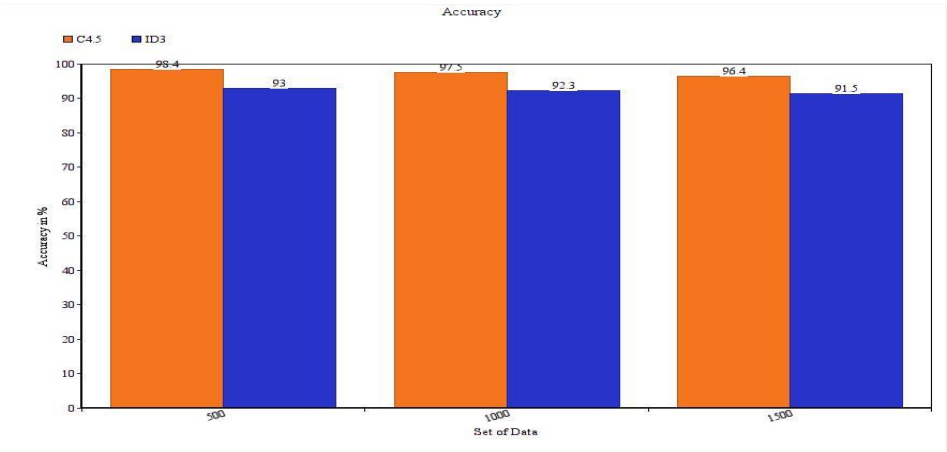

Figure 4: Graph representing Accuracy

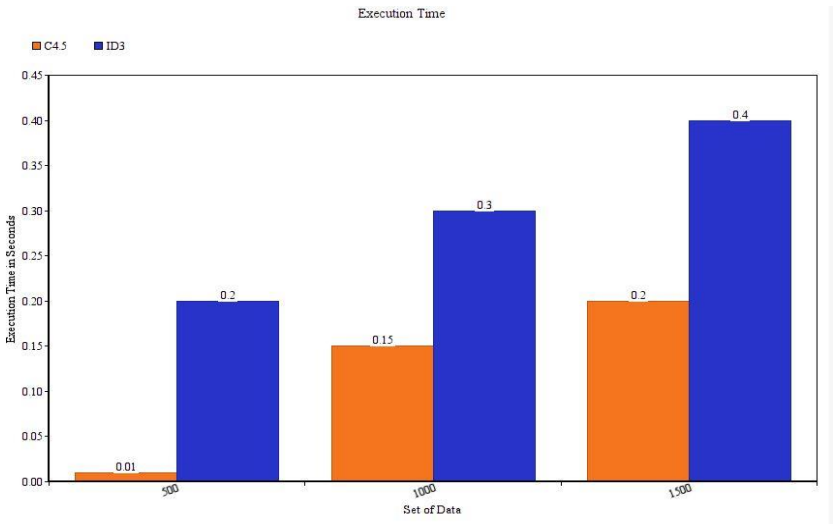

Figure 5: Graph Representing Execution Time

From figure 4 and 5, the accuracy and execution time for C4.5 and ID3 algorithms when considering 500, 1000, 1500 tuples from the dataset can be observed.

ISSN: 0975-8585

CONCLUSION

An Efficient classification framework is achieved for intrusion detection. For experimental purpose KDD99 dataset is used. The Fuzzy values are calculated and assigned as fuzzy state by considering source bytes and destination bytes. By Applying Fuzzy class-association rule mining technique, class value is calculated by considering protocol id, source bytes, destination bytes, fuzzy value and fuzzy state for both intruders and normal users. Certain Rules extracted for identifying intruders and normal users, using judging nodes and processing nodes. The generated protocol id values are sorted and distinctly display the intruder id values and normal user id values. This gives a clear differentiation of anomaly and normal users. C4.5 decision tree provides an efficient classification of intruders with easy representation. The accuracy of normal user and anomaly user detection accuracy is 98.4% with minimal false positive rate. The data reduction model and intrusion detection system with efficient classification framework is achieved.

REFERENCES

[1] Amiri F, Lucas C. Journal of Network and Computer Applications 2011; 34: 1184–1199.
[2] Farid DM, Rahman MZ. Journal of Computers 2010; 5: 23–31.
[3] Frey BJ, Dueck D. Science 2007; 315: 972–976.
[4] Tavallaee M, Bagheri E, Ghorbani A Proceedings of the 2009 IEEE Symposium on Computational Intelligence, Ottawa, Canada 2009; 53-58.
[5] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi and Lilly Suriani International conference on Digital Object Identifier 2010;200-203.
[6] Uppalaiah B, Anand K. IJCST 2012;3:156-160.
[7] Crosbie M,Spafford G. AAAI Fall Symposium 1995;95-01.
[8] Shingo Mabu, CiChen, Nannan Lu, Kaoru Shimada, and Kotaro Hirasawa. IEEE Transactions On Systems, Man, And Cybernetics-Part C: Applications And reviews 2011;41:1.