A Design and Implementation of Intrusion Detection System by using Data Mining

Brijesh Sharma[1] and Huma Gupta[2]
[1] *M.Tech. Scholar CSE Dept. TRUBA Institute of Engg & I.T, Bhopal (M.P.)*
brijesh.sharma2009@hotmail.com
[2] *Dept. Of CSE. TRUBA Institute of Engg & I.T, Bhopal (M.P.)*
huma.g@trubainstitute.ac.in

Abstract: **The role of the intrusion detection system is to enforce the pattern matching policies decided for the network. Basically Proposed IDS executes on the KDD'99 Data set; this data set is used in international level for evaluating/calculating the performance of various intrusion detection systems (IDS). First step is association phase in which frequent item set are produced by apriori algorithm. The second step is clustering phase in which clusters are created by k-Means. Proposed technique uses the standard KDD99 (knowledge Discovery and Data Mining) intrusion detection contest data set. Proposed system can detect the attacks/intrusions and classifies them into different categories: U2R (User to Root), probe, R2L (Remote to Local), and Denial of Service (DoS). The prime task of the proposed IDS is to improve effectiveness with efficiency. An experiment is carried out to evaluate/calculate the performance of the proposed approach using KDD 99' dataset. Here the result shows that the proposed IDS technique performs better in term of efficiency(Execution Speed) & effectiveness.**

Keywords—Internet; Security; intrusion detection; data mining; apriori; Clustering; association analysis; data preprocessing,

I. INTRODUCTION

The goal of this paper is to try to improve the effectiveness of Intruder Detection and to see the possibilities of how the OS Intrusion Detection System might cooperate [2, 3, 4] with Proposed Intrusion Detection System to achieve this goal. The proposed work is to propose a new intrusion detection system approach which is the combination of two technique according to their functionality, without explaining the predefined/fixed intrusion detection system used in that concept. The proposed Intrusion detection system affects the performance of execution and security analysis. In proposed work each issue will be investigated in detail. The proposed approach does rely on specific IDS. The concept of security and the word intrusion detection system might be intimidating and complicated. Proposed Research titled "A Design and Implementation of Signature Based Intrusion Detection System" is an intrusion detection system (IDS) proposed by analyzing the principle of the signature based intrusion detection system based on network. In this performance of the proposed IDS is also evaluated. The proposed IDS approve the effectiveness and efficiency of the proposed technique, and it shows the advantages of network based security. The proposed IDSs offer several advantages like higher security, availability and scalability, over alternative systems and most important thing is that it produced good results in terms of normal and abnormal behaviors of arrived packet. The proposed technique includes integration of individual components to produced good results. It supports a network/system administrator with privileges for finding the intrusions which is reliable, secure and fast. The proposed technique is the combination two data mining technique coming together to produce good results as compared to earlier technique. Proposed work is belongs to signature based intrusion detection system (IDS) and all data mining technique work toward signature based IDS. It's already known that all most all computers/systems are suffering from lower security in the public network as intrusion attacks. Public network are tremendously growing and per day lots of system are connecting in this network and its true that there is no perfect solution to protect computer system from intrusion, but it is very significant to detect them at the initial moment of happening and take necessary actions for reducing the likely damage. There are so many method to handle such type of flaws in this field but intrusion detection system is the an excellent and effective approach to handle intrusion which occurs from the public network [11, 12]. For IDS, various techniques have been applied specifically like artificial intelligence technique, data mining techniques, soft computing techniques, and neural network techniques. In association data mining, proposed technique used modified apriori and in clustering data mining its used K-Mean. The proposed IDS implemented in short time and at a low cost. It also provides a best user interface.

II. PROPOSED WORK

The proposed technique uses data mining techniques. In this apriori and K-mean data mining techniques are applied for signature based intrusion detection field of intrusion detection. Signature based learning technique are able to detect intrusion/ attacks with high accuracy and effectively to gain high detection rates [1, 6, 7]. However, the rate of false alarm using this approach is equally high. To maintain the high accuracy and effectiveness with high efficiency, the proposed technique is combination of two learning techniques. In first approach of the proposed technique, frequent item set of similar data instances are produced based on their behavior by using apriori association technique. In second approach of the proposed technique, similar data instance are grouped, which is based on behaviors of data by using K-Mean clustering to the resulting clusters into attack/intrusion classes as a final clustering task. In this work, *K-Means* as well as *Apriori* algorithm is applied on *KDD* dataset. In order to analyze how these algorithms affect the performance when applied in intrusion detection system[1, 5]. Figure 1 shows proposed approach in which proposed system can be work

978-1-4799-3070-8/14 $31.00 © 2014 IEEE
DOI 10.1109/CSNT.2014.147

on-line as well as off line. For off line data set will be KDD data set for results evolution or to check the performance of the proposed system and for online, proposed system will capture packets which are traveling over network through sensor. In both cases, proposed system sequences the data mining technique like apriori and k-mean data mining algorithms. In this Proposed Technique is designed for off line data set where KDD data set is downloaded and is used to evaluate the performance of the proposed technique.

Following is the proposed IDS which divided into following module**:**

1. Database Selection (Suggested Technique)
   - Data source (KDD 99') is selected
   - Data Pre-processing(Training and Testing)
2. Data mining Techniques
   - Apripori
   - K-Means Cluster Technique
3. Proposed System
   - Apriori Technique
   - Apriori with K-Mean
4. Performance
   - Time Analysis
   - Memory Analysis
   - CPU Analysis

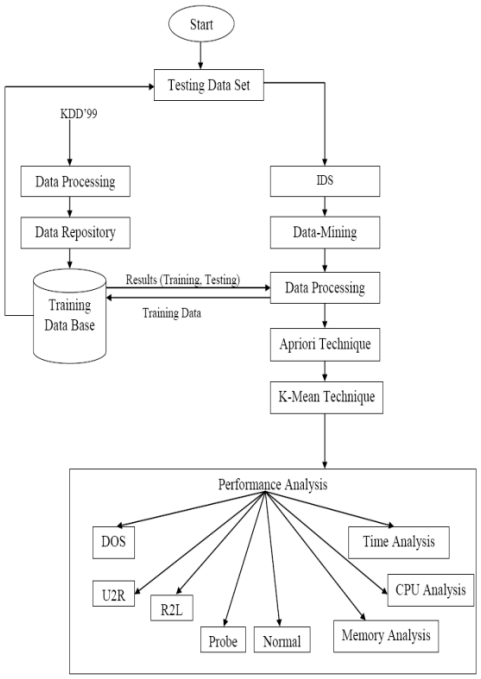


Figure 1 Work Flows of Proposed Technique

***Proposed Architecture*:** In the proposed technique, outline a data mining approach for designing intrusion detection models. The basic idea behind this is that it applies various data mining techniques in single to audit data to compute intrusion detection models, as per the observation of the behavior in the data. The proposed work is the combining two most useable/popular data mining techniques into single technique. In proposed technique, use modified apriori and K-Means (KM) clustering,) algorithm [7]. First apply the modified apriori algorithm to given data set to split the whole data record into frequent item set. Then apply k-means algorithm to the given dataset to split the data records into normal cluster and abnormal clusters. It specifies the number of clusters as five to the k-means and clusters the records in the dataset into normal cluster and abnormal clusters. The abnormal clusters are U2R, R2L, PROBE, and DoS [8, 9, 10]. The records are labeled with the cluster indices. Then, divide the data set into two data sets. One data set is used for training and the other one is used for evaluation. The Proposed work consists of Association and clustering, approach where proposed architecture as shown in figure 2. Then the proposed technique is discussed.

Association Rule Mining: To discover and represent interesting relation among variable in big database, there is a popular technique in data mining is association rule learning. It is indicated to find out strong rules discovered in databases with the help of different interestingness measures [13]. Introduced association rules for finding or discovering products regularities on large-scale transaction data recorded set by point-of-sale systems in supermarkets which are based on the strong rules concept. Apriori algorithm is the best example of the association rule mining technique. Apriori algorithm is a classic algorithm or simple algorithm to mining for frequent item set and association rule learning approach on transactional large databases. It proceeds or continues by identifying or finding the frequent items individually in the large database and increasing them to bigger and bigger item sets and that item sets appeared sufficiently in the database. Apriori algorithm determines the frequent item sets that can be used to identify or determines the association rules to highlight or mark general trends over the record of database: market basket analysis is the best application for such type of domain. Apriori algorithm is designed and developed to operate on large databases which containing transactions. Some other techniques are designed and developed for finding association rules in large data having no timestamps, or having no transactions. Every transaction has a set of items with the threshold value 'TH' [13]. The Apriori algorithm identifies the item sets which are subsets of at least 'TH' transactions in the large database. Bottom up technique or approach is used by apriori algorithm, where at a time only one item extended in frequent subsets (a step called as candidate generation), and such type of candidates groups are tested against the large data. This algorithm terminates or finished when no more successful extensions are identify or found. [13]

Clustering: Clustering is an automated process to group related records together. Related records are grouped together on the basis of having similar values for attributes. Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from

(or unrelated to) the objects in other groups. This approach of segmenting the database via clustering analysis is often used as an exploratory technique because it is not necessary for the end-user/analyst to specify ahead of time how records should be related together. In fact, the objective of the analysis is often to discover segments or clusters, and then examine the attributes and values that define the clusters or segments. k-means clustering is a method of vector quantization originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster [14].

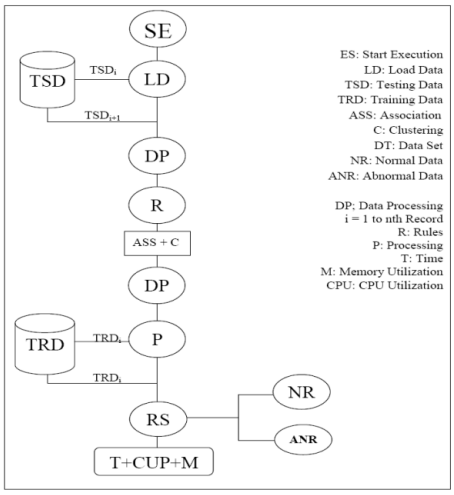


Figure 2: Architecture of the Proposed IDS

Proposed Algorithm:
Input: Dataset KDD, a sample K, Normal Cluster NC, Abnormal cluster AC
Output: **K** is abnormal or normal
Algorithm Hybrid

A) First Apply Apriori

Pass 1

1. Generate the candidate item sets in C_1

2. Save the frequent item sets in L_1

Pass k

1. Generate the candidate item sets in C_k from the frequent item sets in L_{k-1}

2. Join L_{k-1} p with L_{k-1}q, as follows:
   insert into C_k .
   select p.item$_1$, p.item$_2$, . . . , p.item$_{k-1}$, q.item$_{k-1}$
   from L_{k-1} p, L_{k-1}q
   where p.item$_1$ = q.item$_1$, . . . p.item$_{k-2}$ = q.item$_{k-2}$, p.item$_{k-1}$ < q.item$_{k-1}$

   1. Generate all (k-1)-subsets from the candidate item sets in C_k
   2. Prune all candidate item sets from C_k where some (k-1)-subset of the candidate item set is not in the frequent item set L_{k-1}
2. Save the frequent item sets in L_k

B) Second apply K-Mean

Collect data from dataset in the form of frequent item set and apply clustering to the clustering algorithm and build training/testing normal data set D.

1) The dataset is divided into C clusters and the randomly assign data point to the clusters. Generally Clusters and number of data point should be same.
2) For Each data point
   i) Calculate distance between clusters and data points
   If
   Data point Closest Cluster
   Then
   Do not works on that data point leave it where it is.
   Else If
   Data point is not closest cluster,
   Then
   Move it into the closet cluster.
3) Step 2 will continue till every pass completion where all data points should be cover and belong to any cluster resultant there should no data point remain which will move from one of the cluster to another.
4) At that stage stability of data point has formed in the cluster and this clustering process ends.

III. RESULTS ANALYSIS

Proposed design system is the Dot net implementation to present an evaluation system. For timing evaluation of the suggested technique, it is necessary to describe the detailed evaluation method, as illustrated in Figure-1. Here we are taking only one evaluating mode to find Intrusion Detection system impact on time consuming of selected technique. For our experiment use a laptop Pentium® Dual-Core CPU T4400 @2.20Ghz and 32-bit operating system, in which performance data is collected. In the experiments, the laptop executes fixed record data sets (182679) as training data sets and (85131) as testing data sets. Several performance metrics are collected:

- Execution time
- CPU Uses
- Memory Uses

At the time of Results Exploration proposed developed system use the KDD99 data set for training and testing which is shown in table 1 and 2.

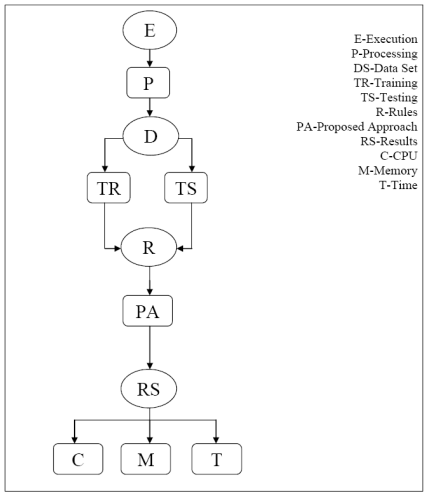


Figure 3: Results Evolution Model

Figure 3 is showing the results evolution model where execution (E) started and perform processing (P) in this processing proposed system configure to training (TR) and testing (TS) data sets (D). once the data set configured then processing called some rules (R) these rules is the combination of modified apriori and K-mean data mining approach known as proposed approach (PA). in this First apply modified apriori algorithm to produced association between attribute then apply K-means clustering algorithm on the features selected. After that, classify the obtained data into Normal or Abnormal clusters and shows the produced results (R) in the form of execution time (T), CPU uses (C) and Memory uses (M). During processing, the record sets are coming from data base. For evaluation mode, there are two parameters: the number of evaluated record set and the size of evaluated record set, where the number of evaluated record sets is the number of record set that are generated randomly and the size of evaluated record sets can be chosen from database. In this mode, n cycles (that is, the number of the evaluated record sets) executed. In each cycle, record sets are respectively executed by existing concept and proposed concept by copying them. The evaluated results are illustrated as in Table 3, 4 and 5.

TABLE 1: Training Data Taken from KDD99 Data Set

Attacks Type	Training Example
Normal	170737
User to Root	245
Denial of service	2065
Remote to User	2331
Probe	7301
Total examples	182679

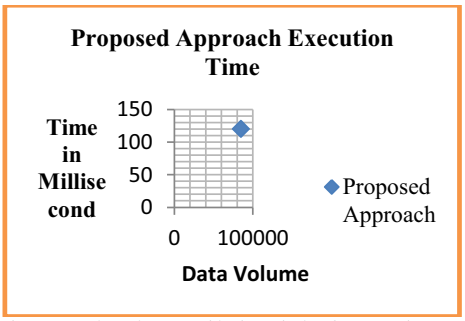
TABLE 2: Testing Data Taken from KDD99 Data Set

Attacks Type	Testing Example
Normal	78932
User to Root	145
Denial of service	885
Remote to User	1015
Probe	4154
Total examples	85131

***Execution Time*:** The execution time considers the time that an approach takes to produce results. Execution time is used to calculate the throughput of an algorithm. It indicates the speed of algorithm. It indicates the speed of approach. Table 3 and graph 1 is showing the execution time of proposed technique on 85131 testing data set.

Table 3: Execution Time Analysis of Proposed Approach

Data Volume	**Proposed Approach**
85131	**Execution Time in Millisecond (Approx)**
	120

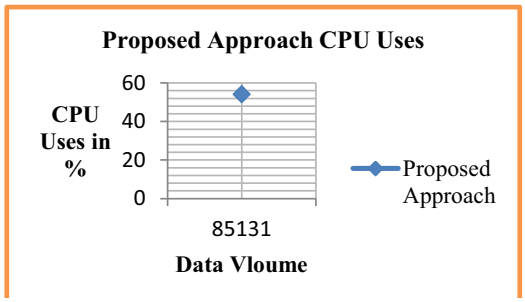


Graph 1: Execution Time Graphical Analysis of Proposed Approach

CPU Uses: - The CPU Uses is the resource used by process during execution time that a CPU is committed only to the particular process of calculations. It reflects the load of the CPU. The more CPU used in the execution process, the higher is the load of the CPU. Table 4 and graph 2 is showing the CPU uses of proposed technique on 85131 testing data set

Table 4: CPU Uses Analysis of Proposed Approach

Data Volume	**Proposed Approach**
85131	**CUP Uses in % (Approx)**
	74

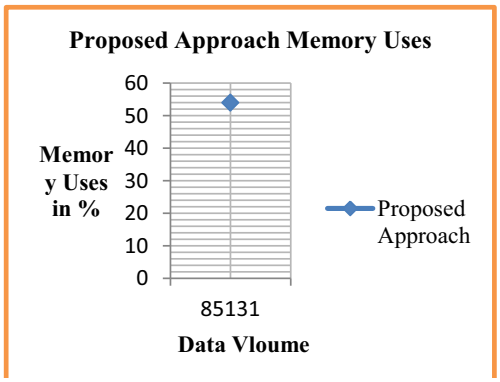


Graph 2: CPU Uses Graphical Analysis of Proposed Approach

Memory Uses: **-** The memory deals with the amount of memory space it takes for the whole process of Intrusion Detection System. Table 5 and graph 3 is showing the memory uses of proposed technique on 85131 testing data set

Table 5: Memory Uses Analysis of Proposed Approach

Data Volume	**Proposed Approach**
85131	**Memory Uses in % (Approx)**
	54



Graph 3: Memory Uses Graphical Analysis of Proposed Approach

Results Analysis: Experimental results for these calculated results are shown Table 3, 4 and 5 at execution stage. The results show the performance of proposed approach in terms of the Execution time, CPU Uses and Memory Uses. Some typical results obtained by the evaluation system can be found in Tables (3, 4 & 5) and Graphs (1, 2 & 3). Finally, it is not difficult to find that, in contrast with these Tables, the larger the data record sets, the bigger execution time is. Besides, in contrast with these Tables, it is not difficult to find that the increasing data length can lead to the significant increment of execution time as well as memory uses and CPU uses.

IV.CONCLUSION

Proposed IDSs offers several advantages over alternative computer/systems. It can be integration of individual components, high availability, scalability, and security. It can provide to admin the privileges for finding the attack/intrusions which is secure, fast and reliable. This type of IDS can be developed and designed in a reasonably short time with very low development cost. It can also provide an attractive GUI's for the users to effective work with. This proposed system can also helps in the network securing. The proposed system can also be extended to find the host-based intrusions with some modification and. The proposed technique is improving the efficiency and effectiveness which is the prime concern of the proposed technique, and presents highly efficient association rule (Modified Apriori) with cluster rules (K-Mean) mining method to abnormal/normal detection experiment based on network. Experimental results proved this method is effective and feasible.

REFERENCES

[1] Chetan R & Ashoka D.V. **"**Data Mining Based Network Intrusion Detection System: A Database Centric Approach" 2012 International Conference on Computer Communication and Informatics (ICCCI -2012), Jan. 10 – 12, 2012, Coimbatore, INDIA.

[2] V. S. Ananthanarayana and V. Pathak "A novel Multi-Threaded K-Means clustering approach for intrusion detection" Software Engineering and Service Science (ICSESS), IEEE 3rd International Conference on 22-24 June 2012 Page(s): 757 - 760 Print ISBN: 978-1-4673-2007-8

[3] N.S Chandolikar and V.D.Nandavadekar, "Efficient algorithm for intrusion attack classification by analyzing KDD Cup 99" Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on 20-22 Sept. 2012 Page(s):1 - 5 ISSN :2151-7681

[4] P.R Subramanian and J.W. Robinson "Alert over the attacks of data packet and detect the intruders" Computing, Electronics and Electrical Technologies (ICCEET), IEEE International Conference on 21-22 March 2012 Page(s):1028 - 1031 Print ISBN:978-1-4673-0211-1

[5] Om, H. and Kundu, A. "A hybrid system for reducing the false alarm rate of anomaly intrusion detection system" Recent Advances in Information Technology (RAIT), 1st IEEE International Conference on 15-17 March 2012 Page(s):131 - 136 Print ISBN:978-1-4577-0694-3.

[6] Virendra Barot and Durga Toshniwal "A New Data Mining Based Hybrid Network Intrusion Detection Model" IEEE 2012.

[7] Wang Pu and Wang Jun-qing "Intrusion Detection System with the Data Mining Technologies" IEEE 2011.

[8] Z. Muda, W. Yassin, M.N. Sulaiman and N.I. Udzir "Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification" 7th IEEE International Conference on IT in Asia (CITA) 2011.

[9] Dewan M.D. Ferid, Nouria Harbi, "Combining Naïve Bayes and Decision Tree for Adaptive Intrusion detection" International Journal of Network Security and application(IJNSA),vol 2, pp.189-196, April 2010.

[10]Dewan M.D. Ferid, Nouria Harbi, "Combining Naïve Bayes and Decision Tree for Adaptive Intrusion detection" International Journal of Network Security and application(IJNSA),vol 2, pp. 189-196, April 2010.

[11] M.Panda, M. Patra, "Ensemble rule based classifiers for detecting network intrusion detection", in Int. Conference on Advances in Recent Technology in Communication and Computing, pp 19- 22,2009.

[12] Joseph Derrick,Richard W. Tibbs, Larry Lee Reynolds "Investigating new approaches to data collection,management and analysis for network intrusion detection".

[13] Sumathi, S.; Sivanandam, S. N.: Introduction to Data Mining and its Applications. Springer, 2006.

[14] LI Min "Application of Data Mining Techniques in Intrusion Detection" 2005