ISSN: 2277- 5528
Impact Factor- 4.015

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & MANAGEMENT

A DATA MINING BASED TECHNIQUE TO ERADICATE THE CHALLENGES BEING FACED WITH THE TRADITIONAL IDS - SIMULATION & RESULT

Mohammed Imroz Malik and Kamal Niwaria

Department of Electronics & Communication Engineering, SRK University, Bhopal (MP) India

e-mail: imroz.19@gmail.com, kamalniwaria@gmail.com,

Corresponding Author: imroz.19@gmail.com

ABSTRACT

The article is going to be present a general idea for intrusion detection system which will enhance efficiency as compare existing intrusion detection system. The proposed concept is using data mining techniques. In this clustering and classification data mining technique has applied for anomaly detection field of intrusion detection. Anomaly learning approaches are able to detect attacks with high accuracy and to achieve high detection rates. However, the rate of false alarm using anomaly approach is equally high. In order to maintain the high accuracy and detection rate while at the same time to lower down the false alarm rate, the proposed technique is the combination of three

INTRODUCTION

This paper is going to be present general idea on a new proposed concept for intrusion detection system which will enhance efficiency as compare existing intrusion detection system. The proposed concept is using data mining techniques. Data mining techniques have been successfully applied in many different fields including marketing, manufacturing, process control, fraud detection, and network management. Over the past five years, a growing number of research techniques have applied data mining to various problems in intrusion detection. In this will apply to data mining for anomaly detection field of intrusion detection. Presently, it is unfeasible for several computer systems to affirm security to network intrusions with computers increasingly getting connected to public accessible networks (e.g., the Internet). In view of the fact that there is no ideal solution to avoid intrusions from event, it is very significant to detect them at the initial moment of happening and take necessary actions for reducing the likely damage. One approach to handle suspicious behaviors inside a network is an intrusion detection system (IDS). For intrusion detection, a wide variety of techniques have been applied specifically, data mining techniques, artificial intelligence technique and soft computing techniques. Most of the data mining techniques like association rule mining, clustering and classification have been applied on intrusion detection, where classification and pattern mining is an important technique.

DATA MINING

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

ISSN: 2277- 5528
Impact Factor- 4.015

DATA MINING WORK

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff, skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data. With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments. While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

DATA MINING CONSISTS OF FIVE MAJOR ELEMENTS

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

ANALYSIS

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square

Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes called the k-nearest neighbor technique.
- Rule induction: The extraction of useful if-then rules from data based on statistical significance.
- Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationship

ISSN: 2277- 5528
Impact Factor- 4.015

ARCHITECTURE

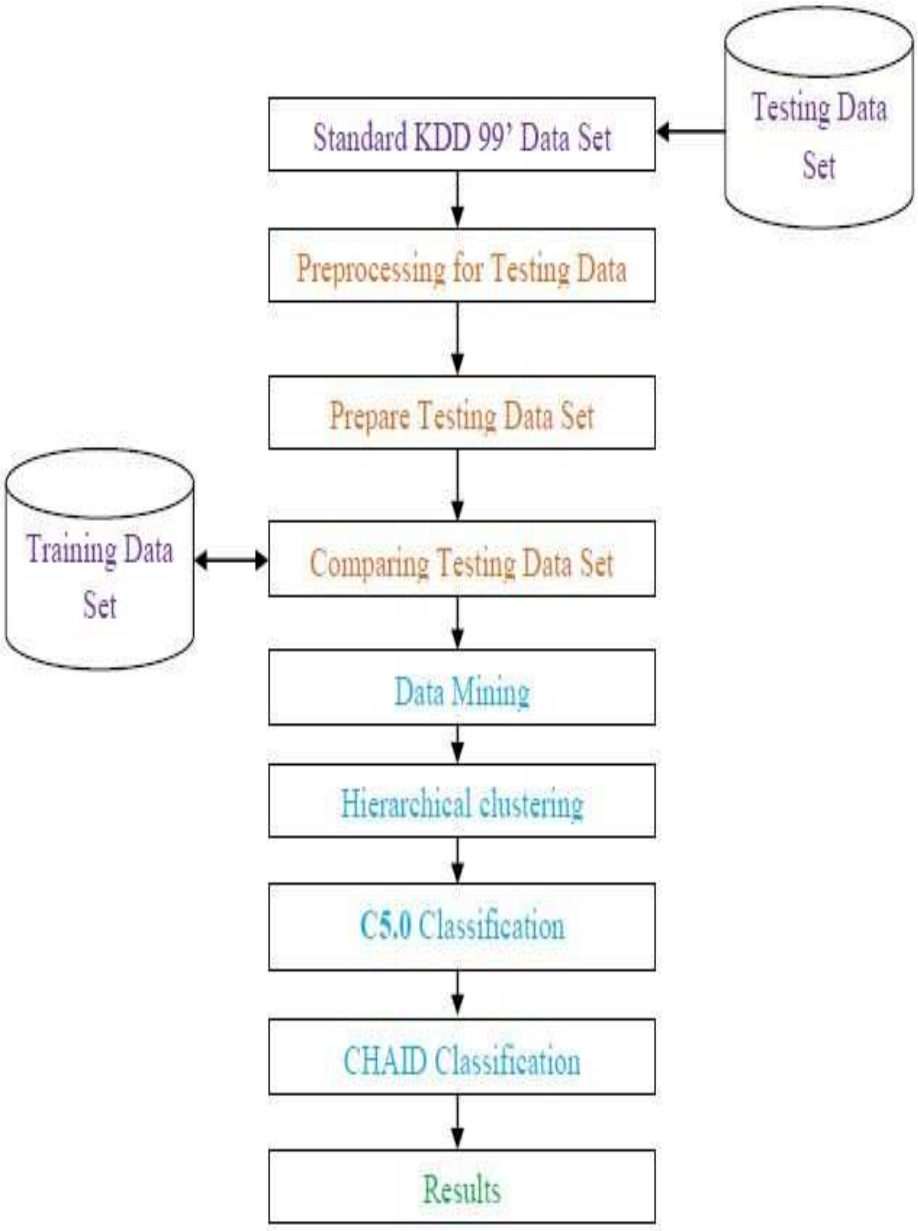

Figure 4.1: Block Diagram

Here we outline a data mining approaches for designing intrusion detection models. The Basic idea behind this is that apply various data mining technique in single to audit data to compute intrusion detection models, as per the observation of the behavior in the data. In the proposed work are the combining three most useable data mining techniques into single concept and presenting architecture shown in figure 4.2. In proposed technique, use Hierarchical clustering, C5.0 algorithm and CHAID approach. First apply the hierarchical algorithm to the given dataset to split the data records into normal cluster and anomalous clusters. It specifies the number of clusters as five to the hierarchical and clusters the records in the dataset into normal cluster and anomalous clusters. The anomalous clusters are U2R, R2L, PROBE, and DoS. The records are labeled with the cluster indices. Then, divide the data set into two parts. One part is used for

training and the other one is used for evaluation. In training phase, apply the labeled records to the C5.0 for training purpose. The C5.0 classifier is trained with the labeled records. Then, apply the rest of unlabeled records to the C5.0 for classification. The C5.0 classifier will classify the unlabelled record into normal and anomalous clusters. Finally apply CHAID which is also the classifier that is doing exact match of each attribute values all to gather and thus removes the strong independence assumption. The Proposed work consists of clustering, classification where proposed architecture as shown in figure 4.2.

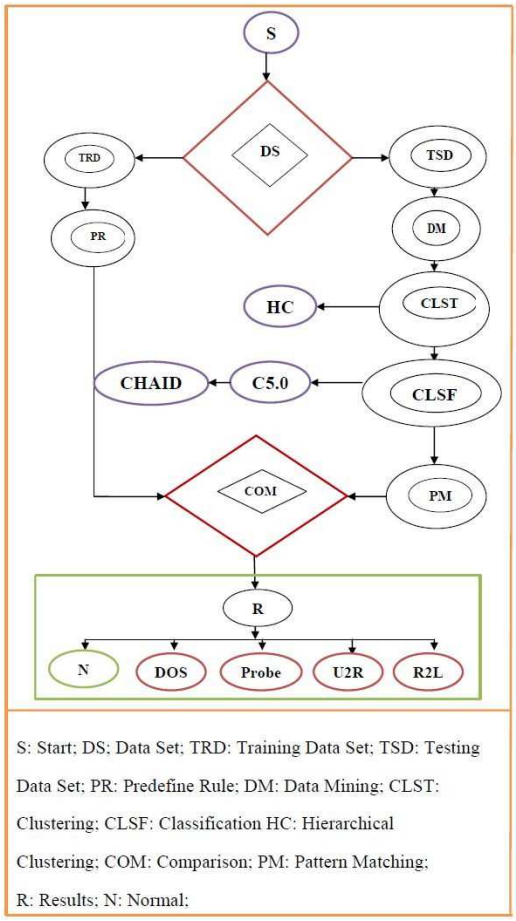

Figure 4.2: Architecture of the IDS

CONCEPT

Here proposed concept are going to be present general idea as showing in figure 4.1 for intrusion detection system which will enhance efficiency as compare existing intrusion detection system. The proposed concept is using data mining techniques. In this clustering and classification data mining technique has applied for anomaly detection field of intrusion detection. Anomaly learning approaches are able to detect attacks with high accuracy and to achieve high detection rates. However, the rate of false alarm using anomaly approach is equally high. In order to maintain the high accuracy and detection rate while at the same time to lower down the false alarm rate, the proposed technique is the combination of three learning techniques. For the first stage in the proposed technique, this grouped similar data instances based on their behaviors by utilizing a hierarchical clustering as a pre-classification component. Next, using C5.0 classifier this classified the resulting clusters into attack classes as a final classification task. This found that data that has been misclassified during the earlier stage may be correctly classified in the subsequent classification stage. At last CHAID classification is applied. Following is the proposed IDS which divided into following module:

1. Database Creation (Suggested Technique)

•Download and Rearranged KDD 99'
•Data Formation and Re-Processing of KDD 99' (Training and Testing Data Set Prepration)

2. Data mining Techniques
   - Cluster Technique
   - o Hierarchical Clustering
   - Classification
   - o C5.0
   - o CHAID

3. Proposed System
   •K-Mean Clustering
   •K-Mean Clustering with Naïve Bayse classification
   •K-Mean with Naïve Bayse classification and Decision Table Majority Rule Based Approach
   - Hierarchical Clustering
   - Hierarchical Clustering with C5.0 classification
   - Hierarchical Clustering with C5.0 classification and CHAID Classification

4. Performance
   •Time Analysis
   •Memory Analysis
   •CUP Analysis

ALGORITHM

Input: Dataset KDD, a sample K, Normal Cluster NC, Abnormal cluster AC, c is the number of clusters and d is the distance between them, ch1,ch2,ch3,ch4,ch5 are Nodes i1,i2,i3,i4 are the category
Output: K is abnormal or normal

Algorithm Hybrid

A) **First apply Hierarchical clustering**

1) Firstly load data into a root cluster and we start with one cluster and successively split clusters to produce others, more and more samples are clustered together in a hierarchical manner.
2) For Every data point:
3) Find out the distance from the data point to every cluster.

Begin
Initialize c; c' = n; Di = {xi}; i = 1,...,n
Do
c' = c' - 1
4) Find nearest clusters Di and Dj

5) Merge Di and Dj
Until c = c'
Return c clusters
End

6) To find the nearest clusters in step 4, the following clustering criterion function is used:
dmin(Di, Dj) = min || x - x' ||, where x € Di and x' € Dj
7) The merging of the two clusters in step 6 simply corresponds to adding an edge between the nearest pair of nodes in Di and Dj. Also, if instead of terminating after a predetermined number of clusters have been obtained; it is possible to set the termination criteria to stop when the distance between nearest clusters exceeds a predetermined threshold.

B) **Apply C5.0 Classification**

1) For each Clusters C in KKDi in test data do
If C is i1
Ch1=c Else

ISSN: 2277- 5528
Impact Factor- 4.015

If C is i2
Ch2=c Else
If C is i3
Ch3=c Else
If C is i4
Ch4=c
Else Ch5=c
until end of data set

2) Collect data from dataset in the form of Normal/Abnormal and apply those data to the CHAID Decision Table Majority rule based approach and build condition for the action like training/testing normal data set D.

C) **CHAID**

3) Preparing predictors. The first step is to create categorical predictors out of any continuous predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of observations. For categorical predictors, the categories (classes) are "naturally" defined.

If (c is not equal to ch1,ch2,ch3,ch4)
Then
c is Normal Otherwise
c is abnormal

SIMULATION AND RESULTS

This IDS system is implemented in GUI is used as a tool for database creation and management. which can be seen as shown in fig

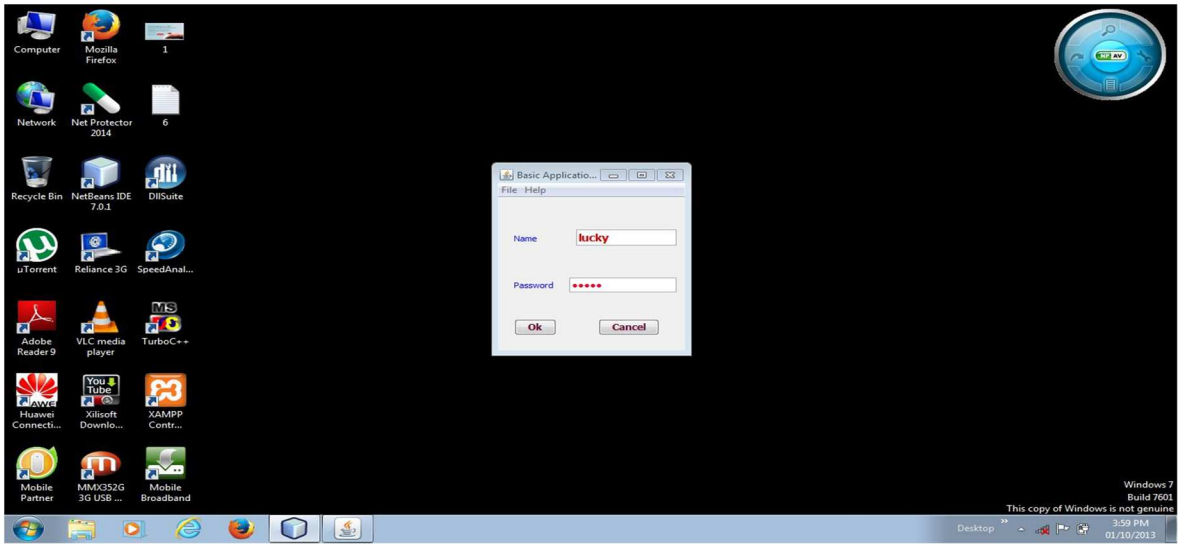

Fig 5.1 First Form "Login Form"

ISSN: 2277- 5528
Impact Factor- 4.015

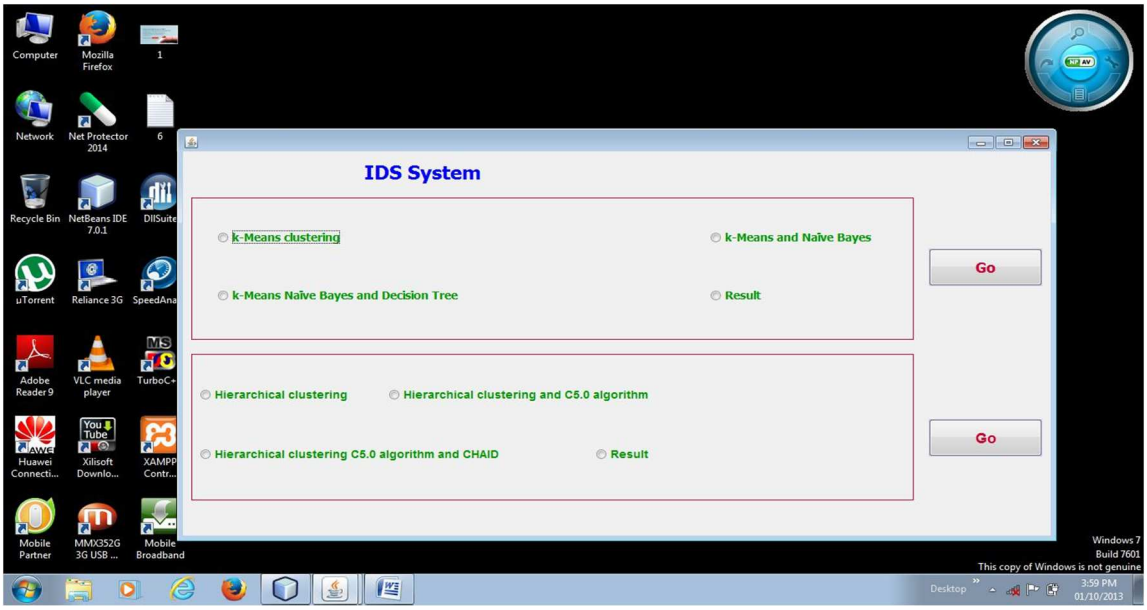

Fig 5.2 Second Form "Selection of Algorithms"

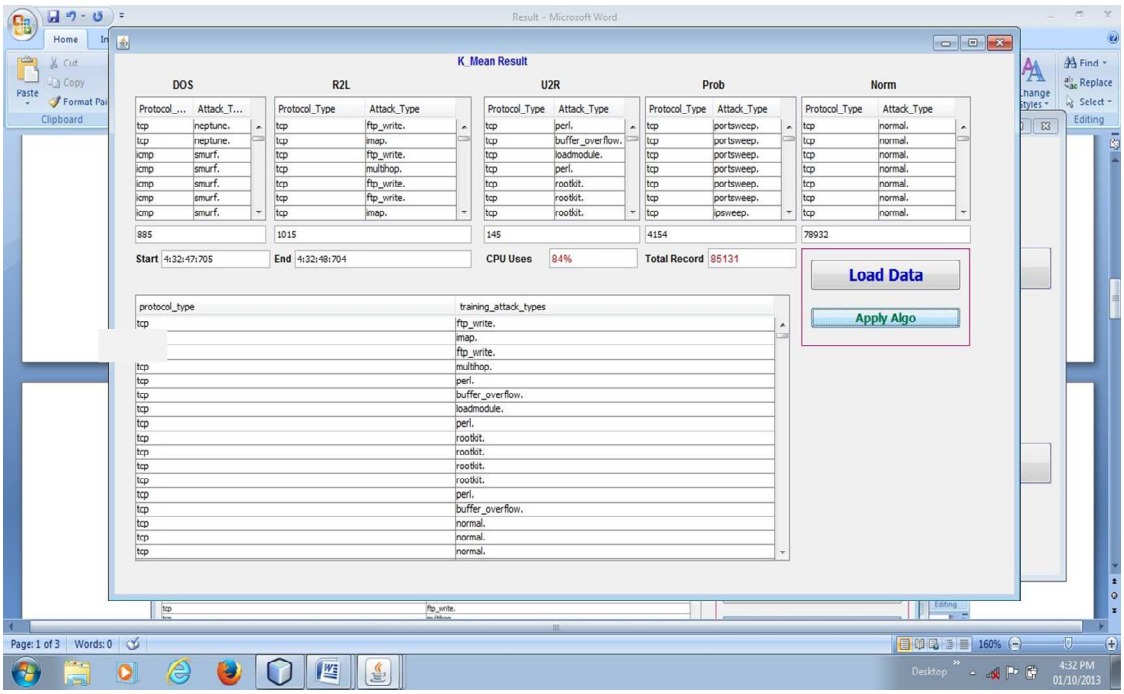

Fig 5.3 Execution of "K-Mean Approach"

ISSN: 2277- 5528
Impact Factor- 4.015

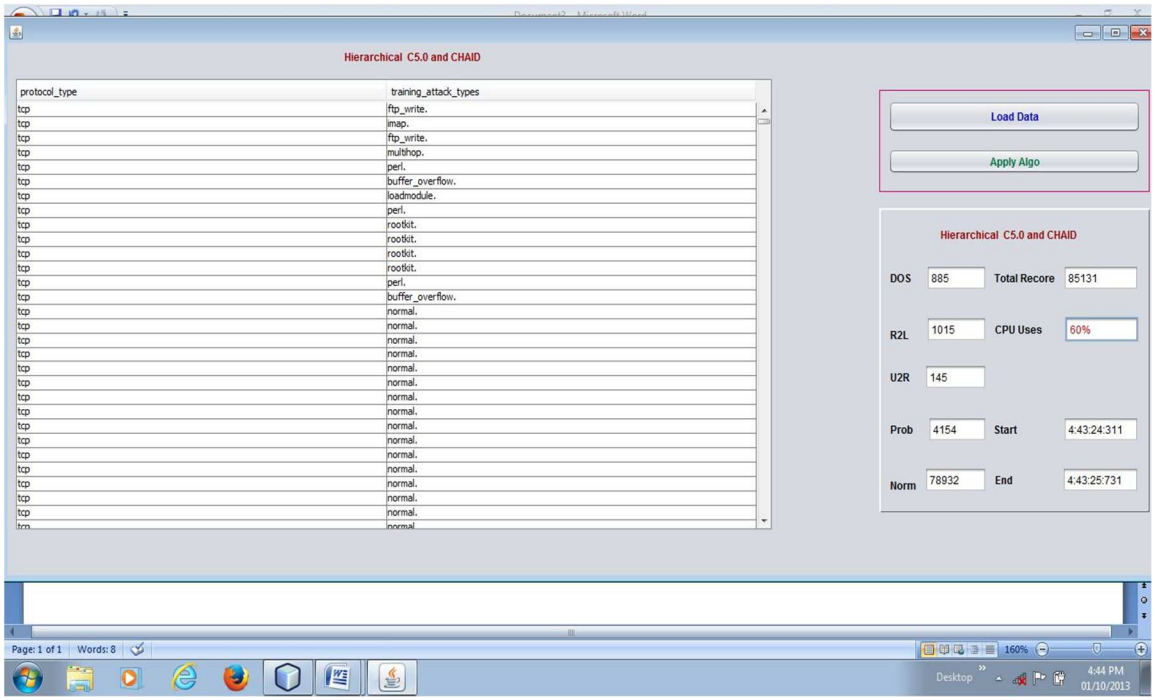

Fig 5.4 Execution of "K-Mean and Naïve Byes" Approach

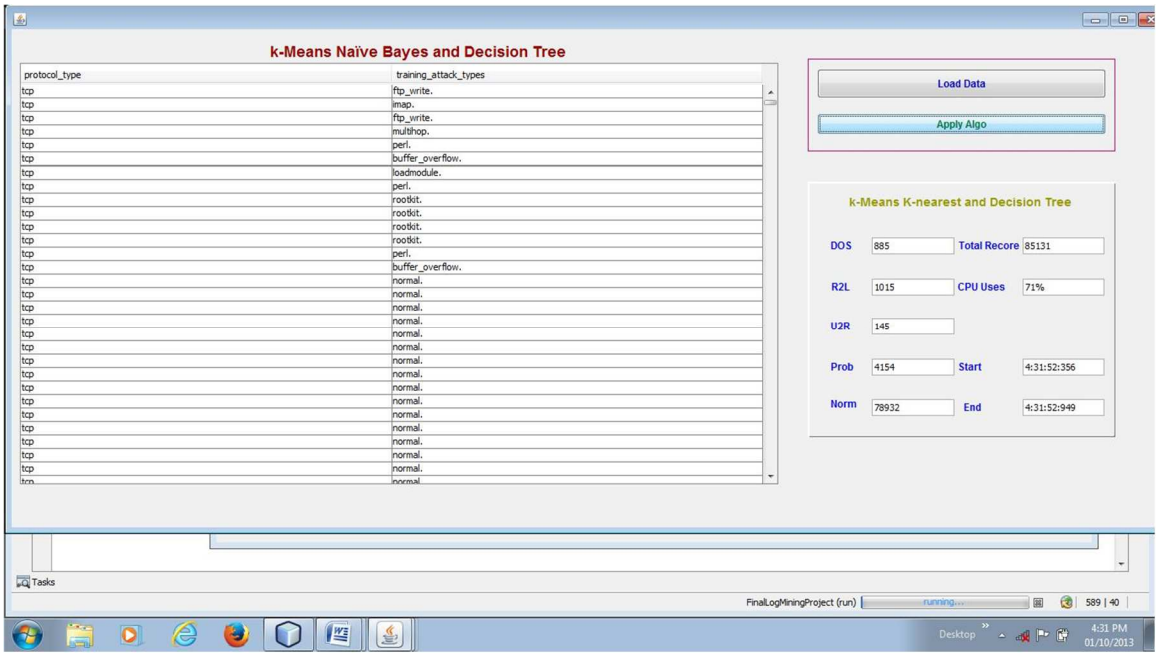

Fig 5.4 Execution of "K-Mean and Naïve Byes and Decision Tree" Approach

ISSN: 2277- 5528
Impact Factor- 4.015

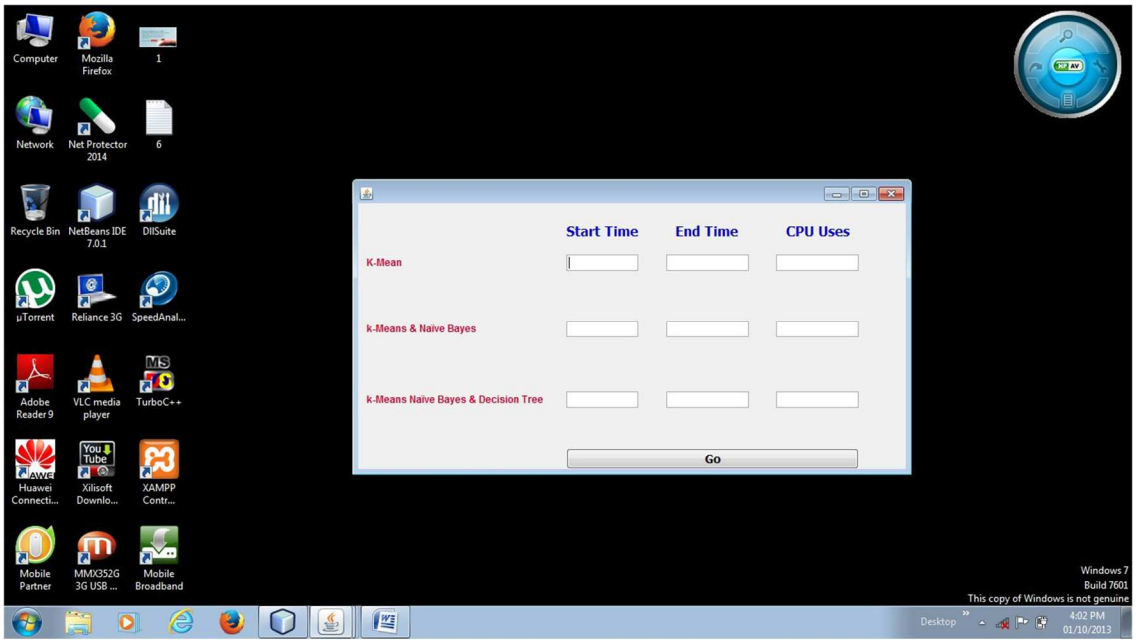

Fig 5.5 Result of Above Three Approach

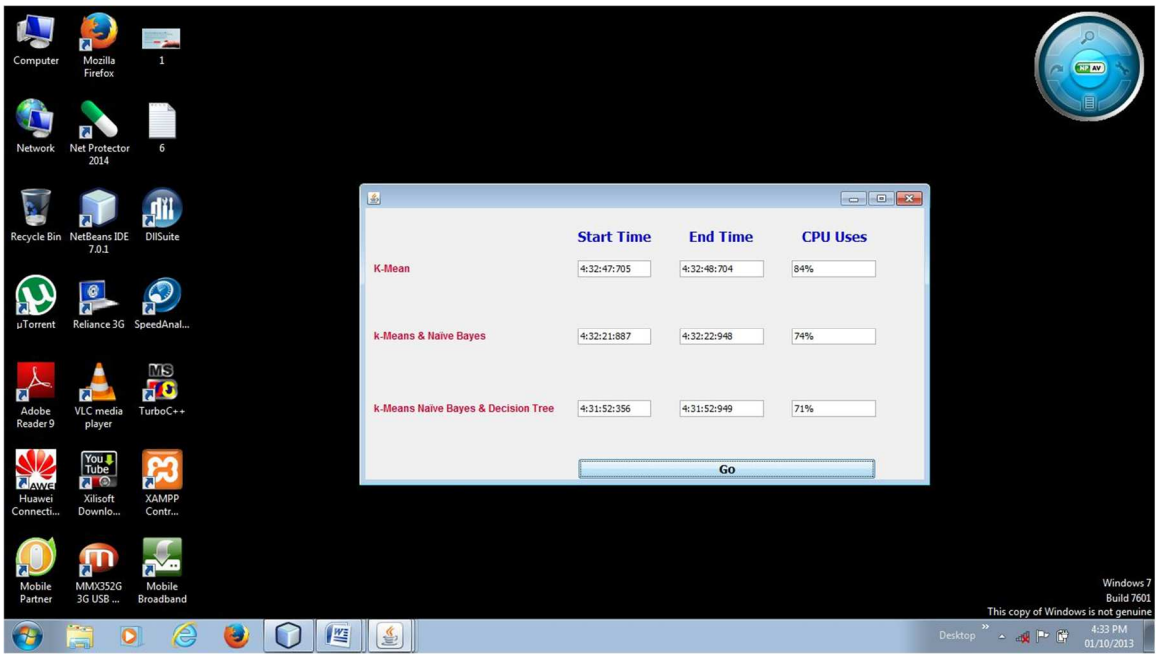

Fig 5. 7 Result of Above Three Approach

ISSN: 2277- 5528
Impact Factor- 4.015

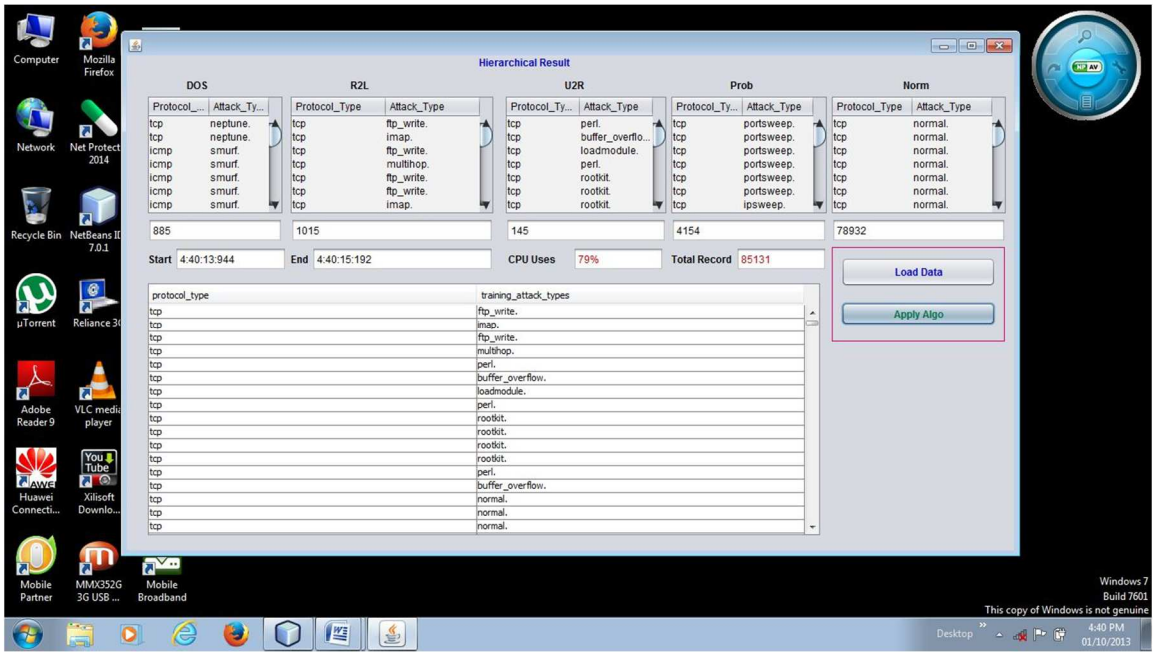

Fig 5.7 Execution of "Hierarchical Clustering" Approach

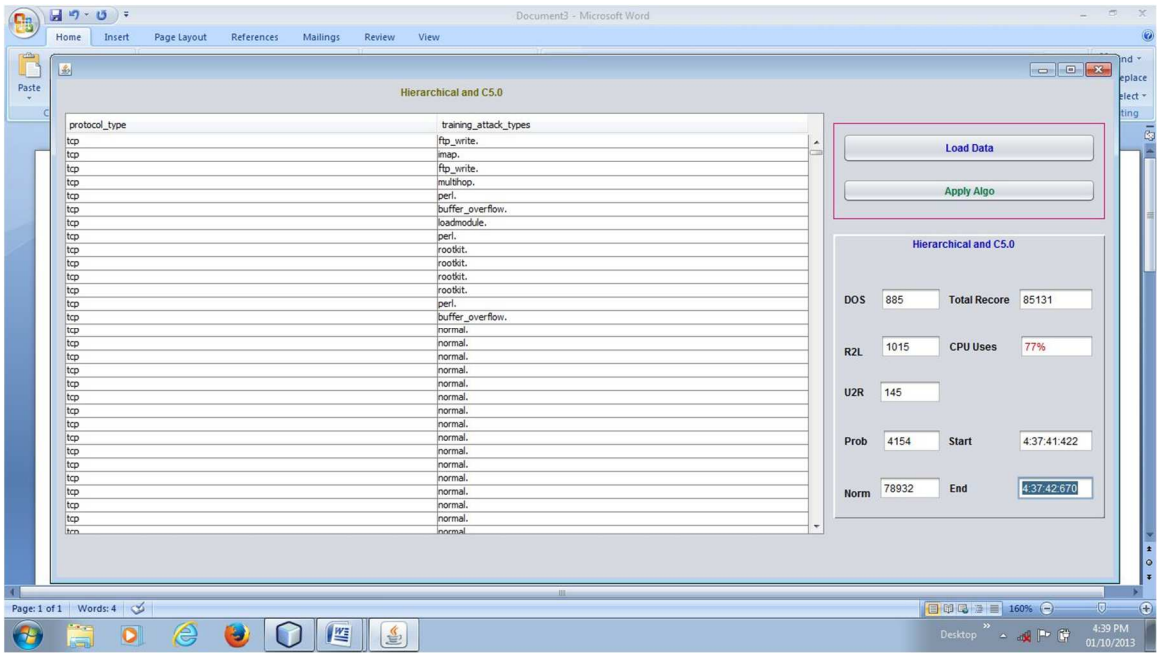

Fig 5.8 Execution of "Hierarchical Clustering and C5.0" Approach

ISSN: 2277- 5528
Impact Factor- 4.015

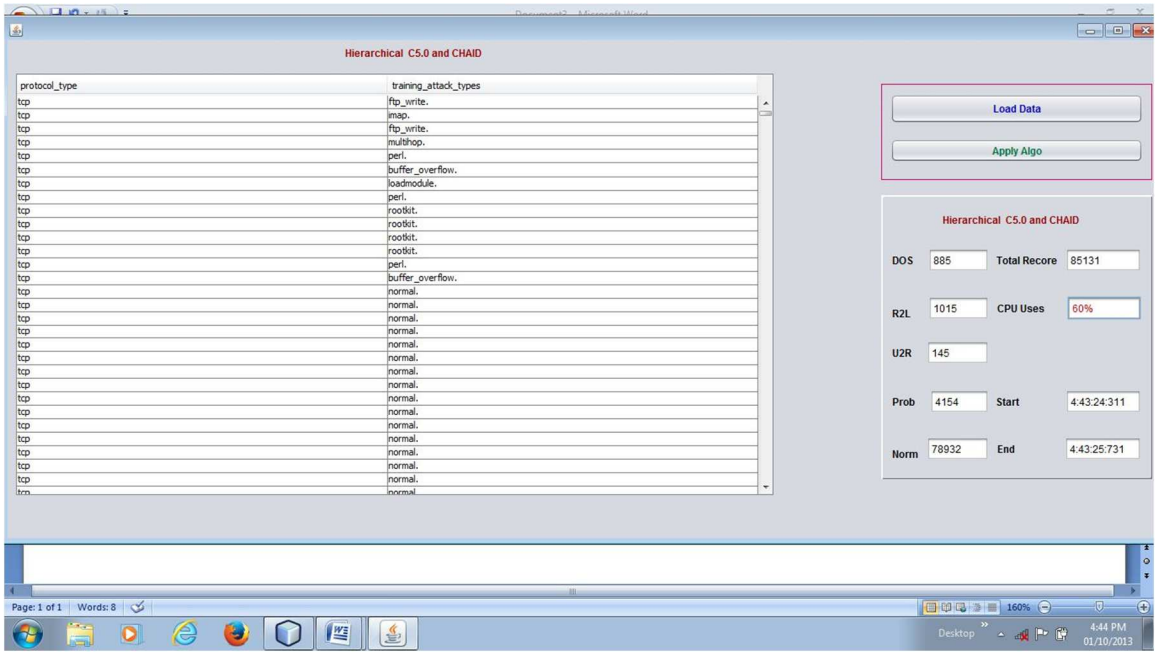

Fig 5.9 Execution of "Hierarchical Clustering, C5.0 and CHAID" Approach

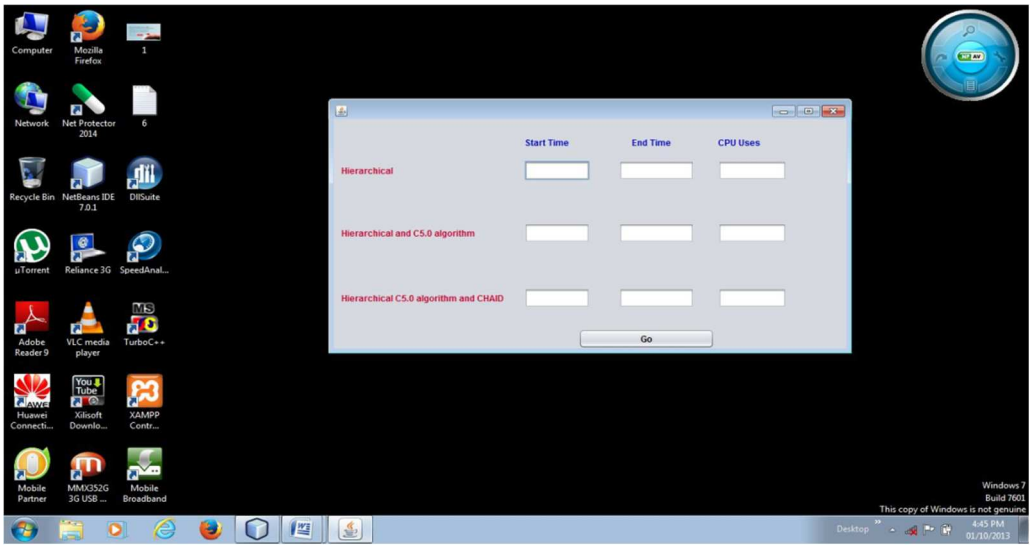

Fig 5.10 Results of Above Three Approach

ISSN: 2277- 5528
Impact Factor- 4.015

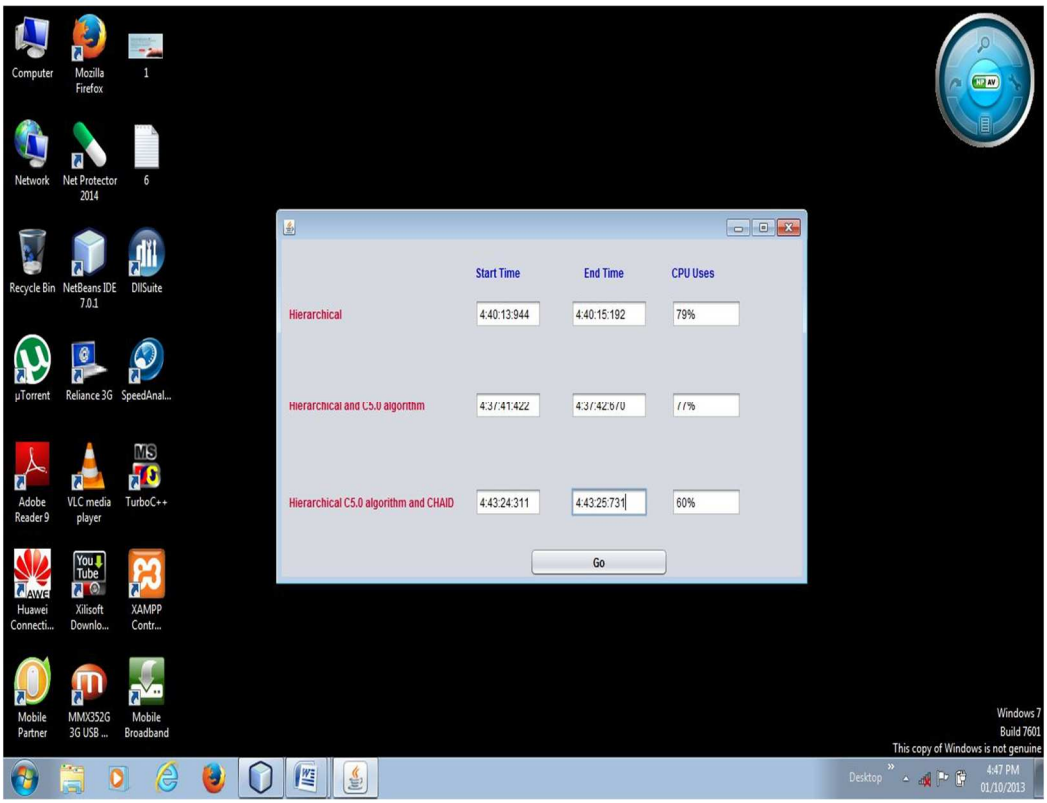

Fig 5.11 Results of Above Three Approach

CONCLUSION

As information systems have become more comprehensive and a higher value asset of organizations, intrusion detection systems has been incorporated as elements of operating systems, although not typically applications. Intrusion detection involves determining that some entity, an intruder, has attempted to gain, or worse, has gained unauthorized access to the system. This research shows that benchmarking intrusion detections systems can be done effectively. In this work design and develop more advanced data mining techniques, it will be very hard to evaluated proposed intrusion detection systems. The amount of customization of data mining techniques that goes into effectively using one, as well as the ever-changing number of viable network exploits makes it impossible at this time. The speed of operation of During Data mining technique is faster. During testing, proposed IDS run in few second to get output. In this no debugging is required. This is due to the high amount of string optimization involved through data mining technique.

REFERENCES

[1] Eric Bloedorn, Alan D. Christiansen, William Hill "Data Mining for Network Intrusion Detection: How to Get Started" 2001.

[2] Skorupka, C., J. Tivel, L. Talbot, D. Debarr, W. Hill, E. Bloedorn, and A. Christiansen 2001. "Surf the Flood: Reducing High-Volume Intrusion Detection Data by Automated Record Aggregation," Proceedings of the SANS 2001 Technical Conference, Baltimore, MD

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

[4] Fayyad, Piatetsky-Shapiro, Smyth: From Data Mining to Knowledge Discovery in Databases. AI MAgazine, 1996.
[5] Roiger, Richard J.; Geatz, Michael W.: Data Mining: A Tutorial- Based Primer. Addison Wesley, 2003
[6] MIT linconin labs, 1999 ACM Conference on Knowledge Discovery and Data Mining (KDD) Cup dataset, http://www.acm.org/sigs/sigkdd/kddcup/index.php?section=1999
[7] The KDD Archive. KDD99 cup dataset, 1999. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[8] M. Tavlle, E. Bagheri, W. Lu, and A. A. Gorbani, "A detailed analysis of the KDD CUP 99 Data Set," Proc. of IEEE Symposium Computational Intelligence for Security and Defense Applications (CISDA'09), pp. 1-6, 2009.
[9] James P. Anderson, "Computer security threat monitoring and surveillance," Technical Report 98-17, James P. Anderson Co., Fort Washington, Pennsylvania, USA, April 1980.
[10] D. E. Denning, "An intrusion detection model," IEEE Transaction on Software Engineering, SE-13(2), 1987, pp. 222-232.
[11] LI Min "Application of Data Mining Techniques in Intrusion Detection" 2005
[12] KDD. (1999). Available at http://kdd.ics.uci.edu/databases/ - kddcup99/kddcup99.html
[13] J. Han, and M. Kamber. Data mining: Concepts and techniques. Morgan Kaufmann. San Francisco, CA. 2001.
[14] Hansen, Hans Robert; Neumann, Gustaf: Wirtschaftsinformatik I. Lucius & Lucius, 2001.
[15] Hipp, Jochen; Guentzer,Ulrich; Nakhaeizadeh, Gholamreza: Algorithms for Association Rule Mining - A General Survey and
[16] http://www.seattlerobotics.org/encoder/mar98/fuz/flindex.html
[17] Northcutt and Novak, Network Intrusion Detection (3rd edition), New Riders, 2003.
[18] Barbará and Jajodia, Applications of Data Mining in Computer Security, Kluwer, 2002.
[19] Han and Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001.
[20] Hand, Mannila and Smyth, Principles of Data Mining, MIT Press, 2001.