RESEARCH ARTICLE OPEN ACCESS

A Design and Implementation of New Hybrid System for Anomaly Intrusion Detection System to Improve Efficiency

Vasim Iqbal Memon*, Gajendra Singh Chandel**
*(M. tech Scholar, Department of Computer Science Engineering, S. S. S. I. S. T, Sehore, RGTU, Bhopal)
** (Professor, Department of Computer Science Engineering, S. S. S. I. S. T, Sehore, RGTU, Bhopal)

ABSTRACT
All most all-existing intrusion detection systems focus on attacks at low-level, and only produced isolated alerts. It is known that existing IDS can't find any type of logical relations among alerts. In addition, they counted very low in accuracy; lots of alerts are false. Proposed research is a combination of three data mining technique to reduce false alarm rate in intrusion detection system that is known a hybrid intrusion detection system (HIDS) combining k-Means (KM), K-nearest neighbor (KNN) and Decision Table Majority (DTM) (rule based) approaches for anomaly detection. Proposed HIDS operates on the KDD-99 Data set; this data set is used worldwide for evaluating the performance of different intrusion detection systems. Initially clusteringperformed via k-Means on KDD99 (knowledge Discovery and Data Mining) intrusion detection after that we apply two-classification techniques; KNN which is followed by DTM. The Proposed system can detect the intrusions and classify them into four categories: R2L (Remote to Local), Denial of Service (DoS), Probe and U2R (User to Root). The prime concern of the proposed concept is to decrease the IDS false alarm rate and increase the accuracy and detection rate.

Keywords-Association Analysis, Clustering,Data Mining, Data Preprocessing,Intrusion Detection

I. INTRODUCTION

With online business more important now than in yesteryears, importance of securing data present on the systems accessible from the Internet is also increasing. If a system is compromised for even a small time, it could lead to huge losses to the organization.

On a daily basis novel tools and techniques are devised to end these malevolent attempts to access or damage information. Conventionally, firewalls have been used to end the intrusion attempts by assailants. But firewalls have inert configurations that obstruct attacks based on a few attributes like destination and source ports with IP addresses. These attributes are not adequate to offer safekeeping from all kind of attacks. Therefore, we need IDS type systems, which could analyze the payload of the packet to detect these attacks [1, 2, 3 & 4]. The motivation of the work is to develop a technique that mediates the user and the operations to achieve security goals with high efficiency.

Users require using the intrusion detection system in tidy to acknowledge attacks in set of connections based system called network. The operations worn cluster of rules to discover the attacks of foreigners to make and read private files that is positioned in own computer or the user would similar to send someplace.

Computers associated in a straight line to the Internet are subject matter to insistent snooping and attack. Whereas shielding events such as modern patching, harmless configuration, and firewalls are all cautious steps they are complex to keep up and cannot assurance that all vulnerabilities are protected. It's known that IDS support defense in depth by detecting and logging hostile activities. An IDS system acts as eye that watches for intrusions as soon as other defensive events fail [7]. Prim concern of the paper to enhance Intruder Detection and to analyze the potential of how the IDS might assist with Proposed IDS to accomplish this. The prime object of the proposed effort is to suggest a new hybrid intrusion detection system thought which is combine three functionality, exclusive of describing the preset intrusion detection system used in that thought. The proposed Hybrid Intrusion detection system affects the executionperformance and analysis of security. The concept of security and the word intrusion detection system might be intimidating and complicated

II. PROPOSED WORK

2.1 Proposed Technique

This section 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 has been fruitfully applied in many diverse fields with manufacturing,marketing, fraud detection, process control, 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 we will apply data mining for anomaly detection field of intrusion detection. Currently, it is unable to be realized for different systems to assert security for network intrusions with system more and more getting connected via Internet. View fact is that there is no perfect approach to avoid or protect intrusions from various events, it is very important to detect or identify them at the initial level of occurrence and take necessary or required actions for reducing or decreasing the likely damage. One move toward to handle doubtful behaviors within a network is an IDS. For intrusion detection, lots of techniques have been functional specifically, soft computing techniques,artificial intelligence technique anddata mining technique. Most of the data mining techniques like, clustering, association rule mining and classification have been functional on intrusion detection, where pattern mining and classification is the significant technique.

2.2 Proposed Concept

Here proposed concept is going to be present general idea as showing in figure 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 K-mean data mining technique has applied for anomaly detection field of intrusion detection. Anomaly learning technique is capable to identifyharms with high correctness and to get large detection rates. On the other hand, false alarm rate using anomaly technique equally soaring. In order to maintain the soaring detection rate and accuracy even as at the same time to decrease the false alarm rate, the proposed technique is the combination of three learning approaches.

For the first stage in the proposed technique, this grouped similar data instances based on their behaviors by utilizing a K-mean clustering as a pre-classification component. Next, using K-nearest-neighbor classifier techniqueit classified the consequential clusters into assails classes as a concluding classification assignment. This found that data that has been misclassified throughout the previous phase might be appropriately classified in the consequent classification phase. At last Decision Table Majority rule based approach applied. Following is the proposed IDS, which divided into following module**:**

a) Database Creation (Suggested Technique)
   – Selecting and generating the data source (KDD 99')
   – Data scope transformation and pre-processing
b) Data mining Techniques
   – K-Means Cluster Technique
   – K-Nearest Classification
   – Decision Table Majority Rule Base Approach
c) Proposed System
   – K-Mean with K-Nearest Neighbor and Decision Table Majority Rule Based Approach
d) Performance
   – Time Analysis
   – Memory Analysis
   – CPU Analysis

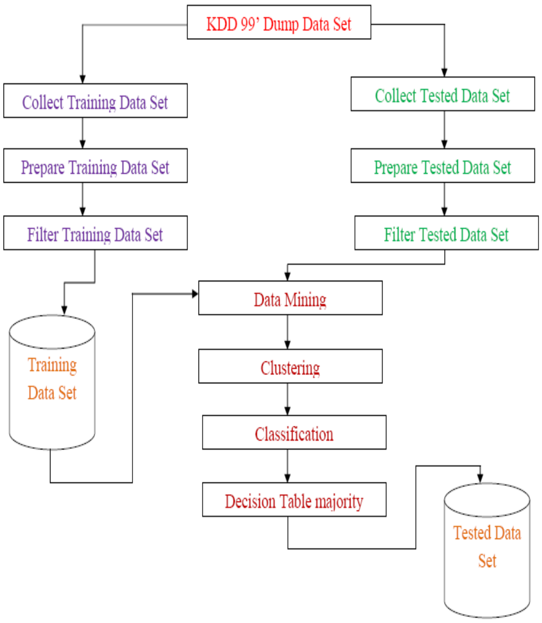


Figure 1: Block Diagram of Proposed Concept

2.3 Proposed Architecture

In the proposed technique, 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 is the combining three most useable data mining techniques into single concept and presenting architecture shown in figure 2.In proposed technique, use K-Means (KM) clustering, K-Nearest Neighbor (KNN) algorithm [7] and Decision Table Majority (DTM) rule based approach. First apply the k-means 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 k-means 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 K-nearest neighbor for training purpose. The K-NN classifier is trained with the labeled records. Then, apply the rest of unlabeled records to the K-Nearest Neighbor for classification. The K-NN classifier will classify the unlabeled record into normal and anomalous clusters. Finally apply decision table majority rule based approach. Decision Table Majority (DTM) is the technique of classifier,which is responsible for correct match of every attribute standards all to meet and thus remove the well-built independence conjecture. The Proposed work consists of clustering, classification and Decision Table majority rule based approach where proposed architecture as shown in figure 2. Then the proposed concept is discussed.

2.3.1 Clustering

Clustering is a division of data into groups of similar kind of objects. Each group or cluster contains objects that are similar among themselves but dissimilar with the others. The greater the difference between groups, the better is the clustering. Clustering is an unsupervised learning because the class labels are not known. A group of measurements and observations are done for the existence of the data in a cluster. Some clustering algorithms are: k-Means [1], Agglomerative Hierarchical clustering and classification and DBSCAN [7]. I use k-means clustering in this work.

2.3.2Classification

This module assigns class labels to the objects. It is trained first with records along with the class labels in the training phase. The data sets are divided into search domain and new samples. It builds a classification model from the search domain and decides the class domain for each given object using one of the methods - k-nearest neighbor [1].

2.3.3Decision Table

Decision Table is one of the possible simplest hypothesis spaces, and usually they are easy to understand. A decision table is a managerial or encoding tool or technique for the demonstration of separate functions. This can be viewed as a matrix where the higher rows identify sets of circumstances and the lesser ones sets of events to be in use while the matching circumstances are fulfilled; thus each column,called a rule, describes a procedure of the type “if conditions, then actions”. Given an unlabelled instance, decision table classifier searches for exact matches in the decision table using only the features in the schema (it is to be noted that there may be many matching instances in the table) [1]. If no instances are found, the majority class of the decision table is returned; otherwise, the majority class of all matching instances is returned. If the training dataset size is, say *D* and test data set size is, say *d* with *N* attributes, The complexity of predicting one instance will be *O(D*N)*. So, the underlying data structure used for bringing down the complexity is Universal Hash table. The time to compute the hash function is *O(n’)* where n’ is the number of features used as schema in decision table. So complexity will become lookup operation for n’ attribute in addition to *l*, number of classes that is *O*(*n’* + *l*). To build a decision table, the induction algorithm must decide which features to include in the schema and which instances to store in the body. More details can be found in [1, 5].

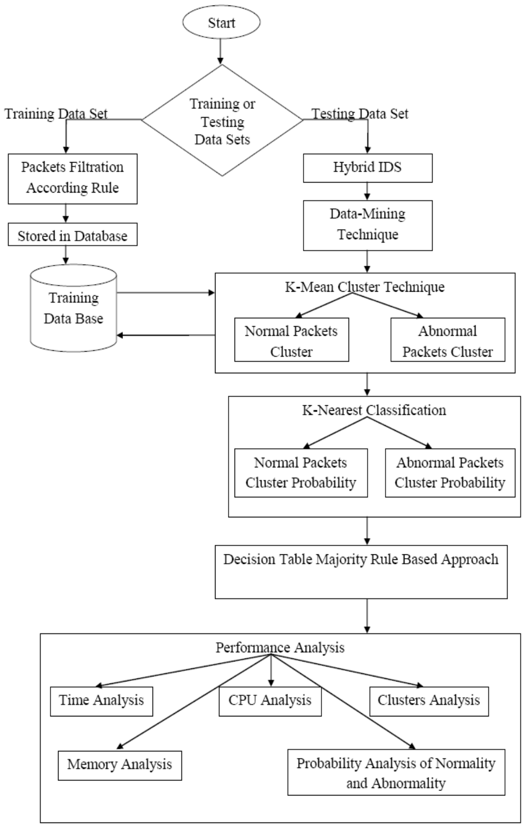


Figure 2: Architecture of the Proposed IDS

2.4 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 K-Means

1) The dataset is divided into N clusters and the data points assigned randomly to the clusters. Roughly Number of data point and cluster are same.

```
2) For Every data point:Find out the distance
   from the data point to every cluster.
   if(Data point  == Nearest Cluster) then
      Leave it where it is
   else if(Data point == is not nearest cluster)
   then
      Move it into the closest cluster
3) Repeat step 2 until pass completion through
   all the data points' resultant there is no data
   point, which is moving from one of the cluster
   to another.
4) At that point stability in the cluster has formed
   and this clustering process ends.
```

Collect data from dataset in the form of Clusters and apply those clusters to the Classification algorithm and build training/testing normal data set D.

B) Apply K-Nearest NeighborClassification

```
1) Collect clusters form KDD of data set, KKD$_i$.
2) For each Clusters K in KKD$_i$ in test data do
   i)  if (Cluster K is in KDD$_i$)then
           K is abnormal
       else
           Find scores of dist ($K_1$, $K_2$), for all $K_1$, $K_2$
           belong KDD$_i$, where $K_2$ is another record
           Cluster or point.
   ii)  Arrange the distances in particular order
        like ascending order or descending here
        ascending order is used.
   iii) Find first k shortest clusters and pick up
        the first shortest k nearest neighbors
   iv)  if ( ($K_1$, N) < ($K_1$, A))then
           $K_1$ is Normal
       else if (($K_1$, N)> ($K_1$, A))then
           $K_1$ is abnormal
```

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

C) Decision Table Majority rule based approach

```
1) Calculate Every Unique record of training
   data set with attribute set S and update counter
   for further prediction process by using DTM.
2) if (Cluster K(Training KDD$_1$) == K(Testing
   KDD$_1$)) then
      K is Normal
   else
      K is Abnormal
```

III. RESULT ANALYSIS

For our experiment use a laptop Pentium® Dual-Core CPU @2.81Ghz and XP operating system. In the presented experiments, the system executes fixed record data sets (182679). Several performance metrics are collected. During Results evolution we used the KDD99 cup data set [20, 21] for training and testing [1] which is shown in table 2 and 3. First apply K-means clustering algorithm on the features selected. After that, classify the obtained data into Normal or Anomalous clusters by using the Hybrid classifier, which is the combination of (K-nearest and Decision Table). During processing, the record sets are coming from database, table 2 is producing training data set and table 3 is producing testing data set. 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 4-7.

Table 1: Attack Classes In KDD'99 Data Set

Four Main Attack Classes	22 Attacks
Denial of Service (DOS)	back, land, neptune, pod, smurf, teardrop
Remote to User (R2L)	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster
User to Root (U2R)	buffer_overflow, loadmodule, perl, rootkit
Probing	ipsweep, nmap, portsweep, satan

Table 2: Number of Example used in Training DataTaken from KDD'99 Data Set

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

Table 3: Number of Example used in Testing DataTaken from KDD'99 Data Set

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

We have applied 10 fold cross validation evaluation on the data set, classification accuracy such as detection rate (DR), false positive rate (FPR), overall classification rate (CR) for evaluating the performance of the intrusion detection task. The meaning of true positive (TP), true negative (TN), false positive (FP), false negative (FN) are defined as follows [1].

– True positive (TP): number of malicious records that are correctly classified as intrusion.
– True negative (TN): number of legitimate records that are not classified as intrusion.
– False positive (FP): number of records that are incorrectly classified as attacks.
– False negative (FN): number of records that are incorrectly classified as legitimate activities.

Detection Rate $=\frac{TP}{TP+FN}$

False Positve Rate $=\frac{FP}{TN+FP}$

Classification Rate $=\frac{TP+TN}{TP+TN+FP+FN}$

Table 4: Result for CPU Utilization

Name	**CPU Utilization in %**(Approx)
Proposed Hybrid Technique (K-Means+K-NN+Decision Table)	**49%**
Existing Technique (K-Means)	**60%**

Table 5: Result for Accuracy

Name	**Accuracy** (Approx)
Proposed Hybrid Technique (K-Means+K-NN+Decision Table)	**96.55%**
Existing Technique (K-Means)	**92.30%**

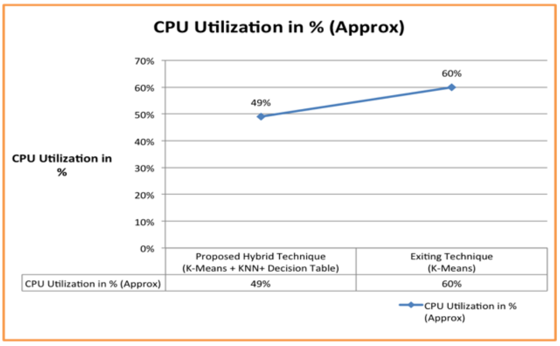
Table 6: Result for Detection Rate

Name	**Detection Rate** (Approx)
Proposed Hybrid Technique (K-Means+K-NN+Decision Table)	**93.67%**
Existing Technique (K-Means)	**91.58%**

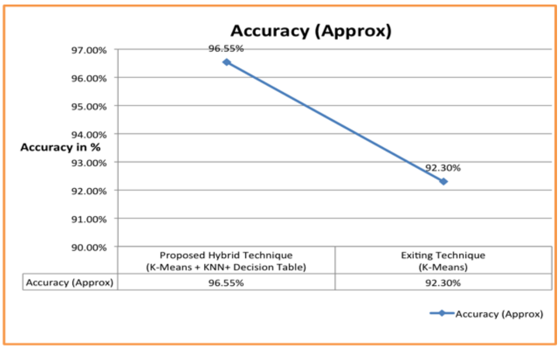
Table 7: Result for False Positive Rate

Name	**False Possitive Rate**(Approx)
Proposed Hybrid Technique (K-Means+K-NN+Decision Table)	**0.019**
Existing Technique (K-Means)	**0.025**

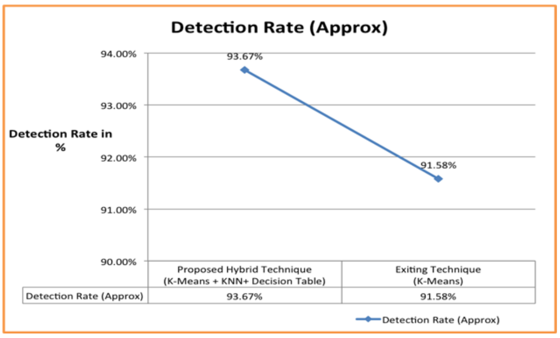
Here, Method 1:kMeans clustering, kNN and Decision Tree Table, Method 2: kMeans clustering. Table 1 shows attack classes in KDD Cup 99 Data set, table 2 and 3 shows number of examples used in the training and testing. The attacks can be divided into 4 major categories, DoS, U2R, R2L, and Probe.



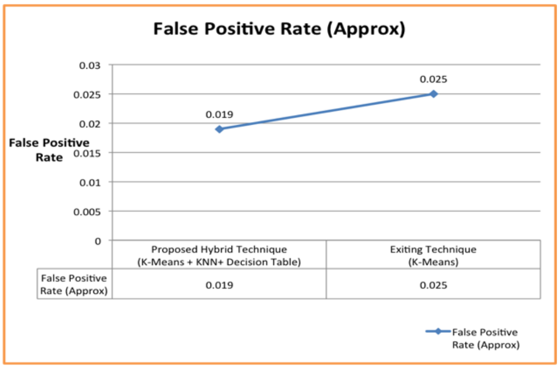
Graph 1: Graphical Representationof CPU Utilization



Graph 2: Graphical Representation of Accuracy



Graph 3: Graphical Representation of Detection Rate



Graph 4: Graphical Representation of False Positive Rate

The CPU Utilization, accuracy, detection rate and false positive rate, are calculated from the confusion matrix table using the given formula and results are given in table 4 to 7. From tables 4-7 and graph 1 to 4, it can see that there is a sharp increase in detection rate, accuracy and decrease in false alarm rate. In Method 2, the detection rate is 91.58%, the false alarm rate decreases to 0.025%, and accuracy decreases to 92.30%. But in method 1, which is a combination of kMeans, kNN and Decision Tree classifier, the detection rate reaches 93.67% and the false positive rate has decreased 0.0192% and accuracy increasing to 96.55%. This shows that our proposed approach is better than the conventional kMeans and kMeans, kNN and naïve bayes. Graph 1 is showing the utilization of CPU in % and it is very clear from the results that CPU usage of the proposed concept only 49% which much better thanfor method 2i.e. 60%.

3.1 Proposed System Strength

- Proposed Hybrid technique is producing good performance then comparing technique to find normal packet performance.
- Proposed hybrid technique having low response time than comparing technique.
- Proposed hybrid technique using low memory space during execution than the compared technique and easy to understand and implement.
- Proposed hybrid technique used simple structure, control flow is well defined and looping structure is also minimized.

IV. CONCLUSION

The proposed research have improved detecting speed and accuracy which is the prime concern of the proposed work, and presents more efficient cluster rules, mining method with classification method to abnormal detecting experiment based on network. Presented Approach is a hybrid approach, which is the combination of K-mean, clustering, K-nearest and Decision Table Majority rule based approach. The proposed approach was compared and evaluated on KDD’99 dataset.

Considering the dependent relations between alerts, it proposed an improved cluster Algorithm with k-nearest neighbor classification; this hybrid approach can find more accurate probability of normal and abnormal packets. Compared with other method, proposed method can find the probability from the training data as well as testing data with high efficiency. Usually when an attack performed, it is very possible that there exist attack cluster transitions. Based on this it use the cluster sequences to filter false alarms generated by IDS, experimental results proved this method is effective and feasible.

Future research work should pay closer concentration or attention to the data mining process.Either more work should address the (semi-automatic) generation of highquality labeled training data, or the existence of such data should no longer be assumed.To deal with some of the general challenges in data mining, it might be best to develop special-purpose solutions that are tailored to intrusion detection

REFERENCES

[1] 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.

[2] 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

[3] 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

[4] 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

[5] Virendra Barot and Durga Toshniwal “A New Data Mining Based Hybrid Network Intrusion Detection Model” *IEEE 2012.*

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

[7] 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.*

[8] Dewan M.D. Ferid, Nouria Harbi, "Combining Naïve Bayes and Decision Tree for Adaptive Intrusion detection",*Intl Journal of Network Security and Application(IJNSA),Vol-2*, pp. 189-196, April 2010.

[9] Joseph Derrick,Richard W. Tibbs, Larry Lee Reynolds "Investigating new approaches todata collection,management and analysis for network intrusion detection". *In Proceeding of the 45th annual southesast regional conference, 2007*. DOI=*http://dl.acm.org/citation.cfm?doid=1233341.1233392*

[10] 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.

[11] 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," P*roceeding of the SANS 2001 Technical Conference,* Baltimore, MD.

[12] KDD. (1999). Available at-*http://kdd.ics.uc i edu/databases/-kdd cup99/ kddcup99.html*

[13] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, Classification and regression trees, Monterey, CA: Wadsworth & Books/Cole Advanced Boks & Software, 1984.

[14] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu,Christine D. Piatko Ruth Silverman, Angela Y. Wu " A Local Search Approximation Algorithm for k-Means Clustering" July 14, 2003 *Annual ACM Symposium on Computational Geometry.*

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

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

[17] Fayyad, Piatetsky-Shapiro, Smyth: From *Data Mining to Knowledge Discovery in Databases. AI Magazine*, 1996.

[18] Roiger, Richard J.; Geatz, Michael W.: Data Mining: A Tutorial- Based Primer. Addison Wesley, 2003

[19] 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*

[20] The KDD Archive. KDD99 Cup Dataset, 1999. *http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.htm*

[21] 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.

[22] James P. Anderson, "Computer security threat monitoring and surveillance," Technical Report 98-17, James P. Anderson Co., Fort Washington, Pennsylvania, USA, April 1980.