International Journal of Database Theory and Application
Vol.6, No.5 (2013), pp.23-34
http://dx.doi.org/10.14257/ijdta.2013.6.5.03

Analysis of KDD CUP 99 Dataset using Clustering based Data Mining

Mohammad Khubeb Siddiqui and Shams Naahid

College of Computer Engineering and Sciences,
Salman bin Abdulaziz University, Kingdom Saudi Arabia
m.khubeb@sau.edu.sa, s.azamali@sau.edu.sa

Abstract

The KDD Cup 99 dataset has been the point of attraction for many researchers in the field of intrusion detection from the last decade. Many researchers have contributed their efforts to analyze the dataset by different techniques. Analysis can be used in any type of industry that produces and consumes data, of course that includes security. This paper is an analysis of 10% of KDD cup'99 training dataset based on intrusion detection. We have focused on establishing a relationship between the attack types and the protocol used by the hackers, using clustered data. Analysis of data is performed using k-means clustering; we have used the Oracle 10g data miner as a tool for the analysis of dataset and build 1000 clusters to segment the 494,020 records. The investigation revealed many interesting results about the protocols and attack types preferred by the hackers for intruding the networks.

Keyword: *KDD 99 dataset, clustering, k-means, intrusion detection*

1. Introduction

As the human population grew in number, so did the data about them. Businesses and various other fields like medicine and others, needed to analyze this data to understand the requirements of people and enhance their services. Statistics was one way of analyzing the available data and obtain results. But with the growing amount of data and advent of computing in various fields, extracting useful information from this data using various sophisticated mathematical models and statistics became possible. This extraction of useful information from large high dimensional databases came to be known as "Data Mining".

Data mining is the analysis of observational dataset to find unsuspected relationship and to summarize large amounts of data in novel ways that are both understandable and useful to data owner in proactive decision making. Data Mining is now possible due to advances in computer science and machine learning. Data Mining delivers new algorithms that can automatically sift deep into your data at the individual record level to discover patterns, relationships, factors, clusters, associations, profiles, and predictions—that were previously "hidden". Using normal reports, Data mining can produce decisions and create alerts when action is required. Data Mining is being widely used in various fields, such as in business for Customer Relationship Management, Marketing, *etc*, in medicine for laboratory research, clinical trials, pharmacology, etc, in forecasting of weather, traffic, *etc*, in aviation for pilot assistance and in research in the areas of astrophysics, medicine, business, security, *etc*. In order to apply the techniques to information security we needed datasets. We used a commonly applied dataset in information security research: The network intrusion dataset from the KDD archive popularly referred to as the KDD 99 Cup dataset. The KDD 99 Cup consists of 41 attributes that is 10% of original dataset means 500,000 rows.

ISSN: 2005-4270 IJDTA

1.1. KDD CUP 99 Data Set

The KDD training dataset consist of 10% of original dataset that is approximately 494,020 single connection vectors each of which contains 41 features and is labeled with exact one specific attack type *i.e.*, either normal or an attack. Each vector is labeled as either normal or an attack, with exactly one specific attack type. Deviations from 'normal behavior', everything that is not 'normal', are considered attacks. [18] Attacks labeled as normal are records with normal behavior. A smaller version 10% training dataset is also provided for memory constrained machine learning methods. The training dataset has 19.69% normal and 80.31% attack connections. KDD CUP 99 has been most widely used in attacks on network. The simulated attack falls in one of the following four categories [9]:

1. Denial of Service Attack (DOS): In this category the attacker makes some computing or memory resources too busy or too full to handle legitimate request, or deny legitimate users access to machine. DOS contains the attacks: 'neptune', 'back', 'smurf', 'pod', 'land', and 'teardrop'.

2. Users to Root Attack (U2R): In this category the attacker starts out with access to a normal user account on the system and is able to exploit some vulnerability to obtain root access to the system. U2R contains the attacks: 'buffer_overflow', 'loadmodule', 'rootkit' and 'perl'

3. Remote to Local Attack (R2L): In this category the attacker sends packets to machine over a network but who does not have an account on that machine and exploits some vulnerability to gain local access as a user of that machine. R2L contain the attacks: 'warezclient', ' multihop', ' ftp_write', 'imap', 'guess_passwd', 'warezmaster', 'spy' and 'phf'

4. Probing Attack (PROBE): In this category the attacker attempt to gather information about network of computers for the apparent purpose of circumventing its security. PROBE contains the attacks: 'portsweep', 'satan', 'nmap', and 'ipsweep'

The major objectives performed by detecting network intrusion are stated as recognizing rare attack types such as U2R and R2L, increasing the accuracy detection rate for suspicious activity, and improving the efficiency of real-time intrusion detection models. This detects that the training dataset consisted of 494,019 records, among which 97,277 (19.69%) were 'normal', 391,458(79.24%) DOS, 4,107 (0.83%) Probe, 1,126 (0.23%) R2L and 52 (0.01%) U2R attacks. Each record has 41 attributes describing different features and a label assigned to each either as an 'attack' type or as 'normal'.

The protocols that are considered in KDD dataset are TCP, UDP, and ICMP that are explained below:

TCP: TCP stands for "Transmission Control Protocol". TCP is an important protocol of the Internet Protocol Suite at the Transport Layer which is the fourth layer of the OSI model. It is a reliable connection-oriented protocol which implies that data sent from one side is sure to reach the destination in the same order. TCP splits the data into labeled packets and sends them across the network. TCP is used for many protocols such as HTTP and Email Transfer.

UDP: UDP stands for "User Datagram Protocol". It is similar in behavior to TCP except that it is unreliable and connection-less protocol. As the data travels over unreliable media, the data may not reach in the same order, packets may be missing and duplication of packets is possible. This protocol is a transaction-oriented protocol which is useful in situations where delivery of data in certain time is more important than loosing few packets over the network. It is useful in situations where error checking and correction is possible in application level.

ICMP: ICMP stands for "Internet Control Message Protocol". ICMP is basically used for communication between two connected computers. The main purpose of ICMP is to send messages over networked computers. The ICMP redirect the messages and it is used by routers to provide the up-to-date routing information to hosts, which initially have minimal routing information. When a host receives an ICMP redirect message, it will modify its routing table according to the message.

Various researchers have analysed the KDD Cup 99 Dataset using various methods, but K-means clustering algorithm using ODM has not yet been applied. This investigation makes use of this technique to analyse and study the pattern of attack types in relation with the protocols.

The remainder of the paper is organized as follows. Section 2 discusses the related work with regards to analyses of the KDD cup 99 dataset. Section 3 provides the methodology of the work. Results are reported in Section 4 and Conclusions are drawn in Section 5.

2. Related Work

The literature survey reveals many results, [1], the authors proposed a real-time intrusion detection system based on the Self-Organizing Map (SOM); an unsupervised learning technique that is appropriate for anomaly detection in wireless sensor networks. The proposed system was tested using KDD'99 Intrusion Detection Evaluation dataset. The system groups similar connections together based on correlations between features. A connection may be classified as normal or attack. Attacks are classified again based on the type of attack. It took the system 0.5 seconds to decide whether a given input represents a normal behavior or an attack. In [2], a data mining algorithm called Clustering and Classification Algorithm Supervised (CCA-S) was developed for intrusion detection in computer networks. The algorithm is used to learn signature patterns of both normal behaviors and attacks. Compared to anomaly detection techniques, signature recognition techniques always produce true alarms, but cannot detect unknown attacks. The algorithm's scalable and incremental learning results in a better performance than two other decision tree algorithms. In [3], the authors addressed the main drawback of detecting intrusions by means of anomaly (outliers) detection, which is the high rate of false alarms when a behavior that has never been seen before is presented. In their work, they added a new feature to the unknown behaviors before they are considered as attacks, and they claim that the proposed system guarantees a very low ratio of false alarms, making unsupervised clustering for intrusion detection more effective, realistic and feasible. In [4], the authors proposed a genetic programming approach for multi-category pattern classification applied to network intrusion detection. They used genetic programming as a pre-processing step is to reduce the input patterns dimension towards a better inter-classes discrimination, and it was achieved through non-linear transformations on the original data sets. The tests were carried out on the KDD'99 Intrusion Detection Evaluation dataset. In [5], the authors addressed three issues related to deploying data mining-based intrusion detection systems in real time environments: accuracy, efficiency, and usability. To improve accuracy, artificial anomalies are used along with normal and intrusion data to produce more effective anomaly detection models. A multiple-model cost-based approach is used to produce low-cost and high-accuracy detection models. To improve usability, adaptive learning algorithms are used to facilitate model construction and incremental updates. In [6], the authors proposed a light weight intrusion detection scheme that can be deployed in wireless sensor networks. In this scheme, nodes have to keep collaborating with their nearest neighbors to decide whether an attack has been launched. Two types of attacks are considered: black hole attacks and selective forwarding attacks. The paper also provides a set of general principles that intrusion detection systems deployed in warless sensor networks should follow. In [7], the authors addressed the complexity of the intrusion detection datasets, as most of them are complex and contain large number of attributes. Some of these attributes may be redundant or do not have significant contribution for intrusion detection. The aim of this work is to specify effective attributes from the training dataset to build a classifier using data mining algorithms. Experimental results on KDD'99 intrusion detection dataset show that the proposed approach achieves high classification rates and reduces false positives in such environment with limited computational resources. In [8], the authors focused on the high rate of false positive in intrusion detection associated with the supervised algorithms based systems. In this paper, an efficient data mining algorithm called random forest algorithm is modified and used to build an intrusion detection system. The system was developed using Waikato Environment for Knowledge Analysis (WEKA), which is an open source Java package of machine learning algorithms for data mining tasks. Experiments were conducted on KDD'99 intrusion detection dataset. In [9], the authors conducted a statistical analysis of this dataset a KDD'99 dataset, the most common dataset widely used to evaluate intrusion detection systems, and found some issues that would result in poor systems evaluation. A new dataset (NSL-KDD) has been proposed. This dataset consists of selected records from the original dataset to overcome those shortcomings. In [10], this paper describes Database-centric Architecture for Intrusion Detection (DAID); a system that leverages data mining within the Oracle RDBMS to address the challenges arise when designing and implementing data mining-based intrusion detection systems. DAID offers numerous advantages in terms of scheduling capabilities, alert infrastructure, data analysis tools, security, scalability, and reliability. It is illustrated with an Intrusion Detection Center prototype. In [11], the authors introduced an intrusion detection system based on Adaptive Resonance Theory (ART) and Rough Set theory. The ART was used to create raw clusters that were refined using Rough Set. As a preprocessing stage, symbolic-valued attributes of the dataset were mapped to numerical values. The proposed system was able to detect not only known attacks, but also new unknown attacks. In [12], the authors investigated the relevance of each feature in the KDD'99 intrusion detection dataset. An approach based on information gain was employed to determine the contribution of the 41 features of the dataset to the attack detection. Experimental results show that some features do not have any contribution to intrusion detection. In [13], the authors introduced an intrusion detection approach based on Adaptive Resonance Theory (ART) and Principal Component Analysis (PCA). In this model, PCA is used for feature selection to reduce the computational complexity and training time of ART. Experimental results show that modifications proposed in this approach improved the speed and accuracy of detection. In [14], an intrusion detection system called Unsupervised Neural Net based Intrusion Detector (UNNID) is introduced. The system provides facilities for training, testing, and tuning of unsupervised Adaptive Resonance Theory (ART) neural networks to be used for intrusion detection. Experimental results show that the system can efficiently detect not only known attacks, but also new unknown ones. To mention a few of the attacks Smurf attacks, also known as directed broadcast attacks, and are popular form of DoS packet floods. Smurf attacks rely on directed broadcast to create a flood of traffic for a victim. The attacker sends a ping packet to the broadcast address for some network on the Internet that will accept and respond to directed broadcast messages, known as the Smurf amplifier. The attacker uses a spoofed source address of the victim. If there are 30 hosts connected to the Smurf amplifier, the attacker can cause 30 packets to be sent to the victim by sending a single packet to the Smurf amplifier [20]. Neptune attacks can make memory resources too full for a victim by sending a TCP packet requesting to initiate a TCP session. This packet is part of a three-way handshake that is needed to establish a TCP connection between two hosts. The SYN flag on this packet is set to indicate that a new connection is to be established. This packet includes a spoofed source address, such that the victim is not able to finish the handshake but had allocated an amount of system memory for this connection. After sending many of these packets, the victim eventually runs out of memory resources [19]. IPsweep and Portsweep, as their names suggest, sweep through IP addresses and port numbers for a victim network and host respectively looking for open ports that could potentially be used later in an attack [19]. In [23], author presented a contribution to the network intrusion detection process using Adaptive Resonance Theory (ART1), a type of Artificial Neural Networks (ANN) with binary input unsupervised training. they presented the feature selection using data mining techniques, towards two dimensional dataset reduction that is efficient for the initial and on-going training, and reduce the dataset both vertically and horizontally, numbers of vectors and number of features.

3. Material and Methods

3.1. Data Collection

The dataset was originally created by DARPA and later used in KDD' 99 Cup for benchmarking analysis of attack over the network, is used in our research work [15]. The training data was processed to about five million connection records. In Oracle 10g the database is prepared and KDD dataset is imported into the designed schema with the help of SQL*Loader tool. The oracle database has an additional benefit of data security, high availability, and load support, fast response time and easy to establish the connection between client and server.

3.2. Process of Data

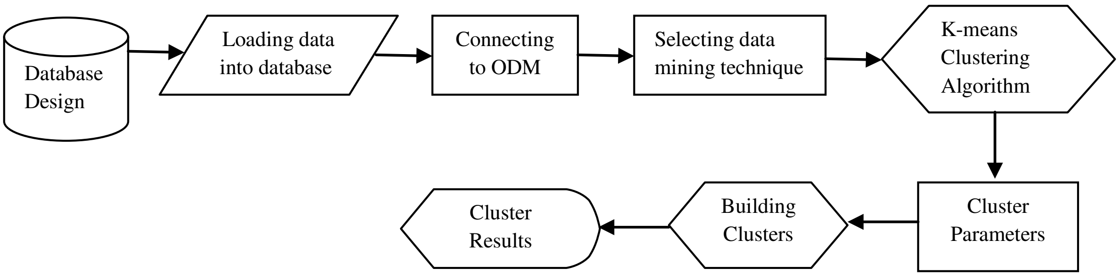
Database Design
Loading data into database
Connecting to ODM
Selecting data mining technique
K-means Clustering Algorithm
Cluster Parameters
Building Clusters
Cluster Results

Figure 1. Modeling of KDD'CUP Data Mining Process

1. Database Design: The database has been designed using Oracle 10g Database. The source table named KDD contains 41 attributes. One primary key attribute has been added.
2. Loading data into database: The 10% of KDD 99 training dataset is a huge dataset having 494,020 records. The available dataset is in .txt format which was imported into Oracle Data Engine using SQL Loader.
3. Connecting to ODM: Oracle Data Miner 10g acts as a client to connect the ODM to Server. It requires certain system privileges from Server side; the system privilege utilized is ctxsys.ctx_ddl.
4. Selecting Data Mining Technique: We had adopted Clustering based Data Mining Technique to design the data model of KDD Cup Dataset.
5. K-Means Clustering Algorithm: Clustering can be done in two ways. The first is K-Means and the second O-Means Algorithm. K-Means Algorithm is used in data pre-

processing steps to identify homogenous group. Here we have used K-Means distance based algorithm with specific number of clusters.

6. Cluster Parameters: Deciding on the values of the following parameters number of cluster (k), distance function, split criterion, Maximum iterations, Number of Bins, Minimum Error Tolerance, minimum support, Block Growth.
7. Building Clusters: The clusters had been processed using ODM on the basis of desired cluster parameters.
8. Cluster Results: The results are displayed in the form of clusters along with the centroid value for each attribute.

3.3. Tools and Techniques

Different types of data mining tools are available and each has its own merits and demerits, for the analysis of 10% of KDD 99 training dataset, present work concentrated on which attack is more common over the network using k-means clustering technique of data mining. The Oracle data miner (ODM) version 10.2.0.3.0; build 2007 is applied for the mining activity, it acts as a client and oracle 10g database releases 10.2.0.3.0 as a server [22].

Oracle 10g Data Miner (ODM) is the mining tool that had applied for the present research work. Oracle Data Miner supports supervised learning techniques (classification, regression, and prediction problems), unsupervised learning techniques (clustering technique and feature selection problem), and attribute importance technique. The availability of these algorithms provides all the necessary tools required in gathering information from dataset. The main advantage of using Oracle Data Miner is that all data mining processing occurs within the oracle database.

Table 1. Type of Attacks Grouped by Protocol

Protocol_Type	Attack_name
UDP	normal, teardrop, satan, nmap, rootkit
TCP	normal, neptune, guess_passwd, land, portsweep, buffer_overflow, phf, warezmaster, ipsweep, multihop, warezclient, perl, back, ftp_write, loadmodule, satan, spy, imap, rootkit
ICMP	normal, portsweep, ipsweep, smurf, satan, pod, nmap

The 10% of KDD 99 training dataset has three distinct protocols namely TCP, UDP, and ICMP. Study reveals that these protocols are interrelated to any of the network attacks. Table 1 show the list of attacks on each of the protocol type which was retrieved from the KDD 99 cup dataset. We have designed a mathematical set of equations (1), (2) and (3) which state the number of attacks in protocol type 'TCP', 'UDP' and 'ICMP' respectively.

$A = \{x|\ x\in \text{attack\_name}\ \forall\ \text{TCP}\in \text{protocol\_type}\}$ ---- (1)
$B = \{y|\ y\in \text{attack\_name}\ \forall\ \text{UDP}\in \text{protocol\_type}\}$ ---- (2)
$C = \{z|z\in \text{attack\_name}\ \forall\ \text{ICMP}\in \text{protocol\_type}\}$ ---- (3)

The following results were obtained:
Equation (1) -‘TCP’ protocol got affected by twenty different attack types
Equation (2) - ‘UDP’ protocol got affected by five different attack types
Equation (3) - ‘ICMP’ protocol got affected by seven different attack types

From the above 3 equations we obtain equation (4) which shows the intersection of equation 1, 2, and 3 applied to the training dataset, wherein the three attacks, normal, satan, and nmap are found to be common in all the three protocols, namely, TCP, UDP, ICMP.

$A \cap B \cap C = \{a | a \in A \text{ and } a \in B \text{ and } a \in C\}$ ---- (4)

In the 10% KDD cup training Dataset the attribute attack_name contains a value either normal or an attack out of the above mentioned 22 attacks. All of the records in the Dataset are divided into 2 classes: Normal or Attack (anomaly) normal class is a defined behavior for the dataset [9]. Any deviation from this Normal behavior is said to be an attack. Again each of these attacks has a particular definition.

3.4. Clustering

Clustering or Cluster Analysis as it is widely known is a focused type of data mining technique for large scale analysis of datasets. Cluster Analysis is a pattern discovery procedure, whose goal is to discover patterns in a set of data. It identifies clusters in a set of data and builds a typology of sets using a certain set of data. In the present research analysis clustering technique is applied it is well known unsupervised data mining technique. It is particularly useful where there are many cases and no obvious natural grouping. Here, clustering data mining algorithm can be used to find whatever grouping may exist. A cluster is a collection of data objects that are similar in some sense to one another. A good clustering method produces high quality cluster to insure that the inter-cluster similarity is low and the intra-cluster similarity is high; in other words, members of a cluster are more like to each other than they are like members of different clusters.

3.4.1. Algorithm for Clustering: ODM provides two type of algorithm for clustering, K-means Clustering and O- means clustering. In the present investigation k-means clustering had been applied it's a distance based clustering algorithm. The pseudo code of k-means clustering given by [21] is written below:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

3.4.2. Centroid: The centroid represents the most typical case in a cluster. The centroid is a prototype. It does not necessarily describe any given case assigned to the cluster. The attribute values for the centroid are the mean of the numerical attributes and the mode of the categorical attributes.

3.5. Experimental Analysis

We have carried out the analysis of attack on 10% of KDD 99 training dataset using K-means clustering technique, we had clustered the training dataset which consisted of 494,019 records and made 1000 clusters. The parameters and the values applied for the analysis to execute the K-means clustering technique in ODM are mentioned as follows:

Distance Function: Euclidean function is used as k-means clustering algorithm which creates clusters by measuring the inter-cluster and intra-cluster distances. ODM supports two methods; Euclidean and Cosine. The Euclidean Distance formula for n-dimensional space is given as [17]:

$$d_2(\vec{a} - \vec{b}) = \sqrt{\sum_{k=1}^{n} |a_k - b_k|^2}$$

where d_2 is the distance between vectors $\vec{a}$ and $\vec{b}$; is the sum of squares of difference between the coordinates of each vector in n-dimensional space.

Split criterion: KMNS_VARIANCE

Split criterion takes two values variance and size. The split criterion is related to the initialization of the K-Means clusters. The algorithm builds a binary tree and adds one new cluster at a time. Split by size results in placing the new cluster in the area where the largest current cluster is located. Split by variance places the new cluster in the area of the most spread out cluster. The node with the largest variance is split to increase the size of the tree until the desired number of clusters is reached.

$$s^2 = \frac{\sum(x_i - \bar{x})}{n - 1}$$

where s^2 is the variance given by average of the squared differences from mean of all the values in consideration

Maximum Iteration: 3

The number of times the K-Means algorithm is to be applied to the dataset. The value can be set anywhere from 2 to 20. The default value is 3. The model can be built fast by decreasing the number of iterations but the result would be poorly defined clusters.

Minimum error tolerance: 0.01

The minimum tolerance controls the convergence of the algorithm. The smaller the value of minimum tolerance, the closest is the algorithm to the optimal solution at the cost of longer run times. This parameter interacts with the number of iterations parameter. The range values for Minimum error tolerance are 0.001(slower) to 1(faster).

Number of Bins: 10

Binning also referred to as Discretizing, is the process of grouping related values together. It reduces the number of distinct values for an attribute, thus building a faster model which is also compact. Equi-width binning is the method used in K-Means. All attributes have the same number of bins. Default value is 10.

Minimum Support: 0.1

In order for the attribute to be included in the rule description for the cluster, the fraction for attribute values that must be non-null is set using Minimum support. The range values are >0 and <=1. If the parameter value is set too high in data with missing values then it can result in very short or empty rules.

Block Growth: 2

Block growth value sets the growth factor for memory allocated to hold cluster data. The range values are from 1 to 5. Default value is 2.

Table 2. Number of Attacks Before and After Clustering

Attack Name	Attacks in training dataset	Attacks in 1000 clusters	Category
normal.	97,277	497	NORMAL
neptune.	101,201	184	DOS
portsweep.	1,040	68	PROBE
satan.	1,589	47	PROBE
warezclient.	1,020	15	R2L
back.	2,203	30	DOS
ipsweep.	1,247	26	PROBE
buffer_overflow.	30	22	U2R
multihop.	7	5	R2L
loadmodule.	9	7	U2R
teardrop.	979	14	DOS
rootkit.	10	8	U2R
ftp_write.	8	5	R2L
imap.	12	8	R2L
guess_passwd.	53	9	R2L
nmap.	231	18	PROBE
smurf.	280,790	18	DOS
warezmaster.	20	4	R2L
perl.	3	1	U2R
spy.	2	2	R2L
pod.	264	8	DOS
land.	21	3	DOS
phf.	4	1	R2L

Total number of cases	494020
Total Number of Clusters	1000

This result was fetched from the clustered database 1000 Cluster i.e. kdd_new_cht_1000

P= n(attack_name) ∃protocol_type(TCP) ---- (5)
Q=n(attack_name) ∃protocol_type(UDP) ---- (6)
R=n(attack_name) ∃protocol_type(ICMP) ---- (7)

P indicates the number of attacks on the protocol type TCP. Q indicates the number of attacks on the protocol type UDP. R indicates the number of attacks on the protocol type ICMP. The above set of equations (5), (6) and (7) show the count of centroid values for the attributes attack_name and protocol_type belongs to TCP, UDP and ICMP respectively.

4. Results and Discussion:

After applying the K-Means clustering algorithm using ODM on 10% KDD cup training dataset, quiet interesting results were discovered in the variation of frequency of attacks. Here we concentrate on the attribute Attack_name which contains 22 distinct attacks (anomalies) and normal type.

Table 3. Category wise Attacks on Protocols in Clusters

Protocol Attacks	TCP	UDP	ICMP
DOS	51.42	35	61.90
U2R	8.53	5	0
R2L	11.61	0	0
PROBE	28.44	60	38.09

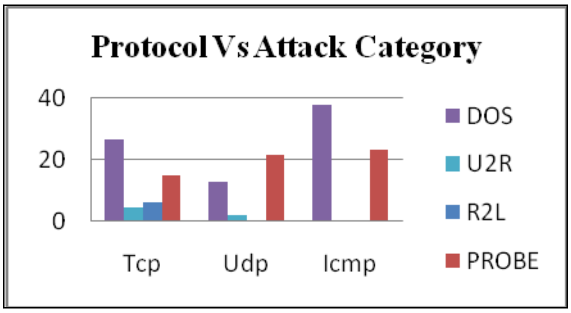

Figure 2. Statistical View of Category Wise Attack on Protocol Types

The training dataset has a maximum of Smurf and Neptune attacks respectively and the normal records accounted to 19% of the total records. Smurf has the maximum frequency with 57% of all the attacks; Neptune is the second most frequent with 21% of the total attacks.

The result shown in Table 3 and its statistical view in Figure 2 obtained after 1000 clusters indicate that 50% of the total clusters have attacks with high count of centroid value. In TCP protocol all the four attack categories DoS, R2L, U2R and PROBE are active as shown in Figure 2. TCP is affected by 19 different attack types in the KDD cup training dataset. Hackers exploit the protocol 'TCP' the most. One reason can be that TCP/IP protocol was designed to be robust. It was designed to recover automatically from any router, line or node failure. Automatic Recovery in TCP is the major reason for the network problems to be undiagnosed and henceforth uncorrected for uncertain period of time. Once a message is sent to an IP Router, it makes an independent decision about where to send it next and which path to choose using routing algorithms. Any problem in any of the routes and the router changes the route automatically and delivers the message to the destination. This architecture of ‘TCP’ has been the main reason for many hackers making DoS attacks and still be unnoticed for long periods of time. Many researches are being carried to find out a solution to this problem.

The hackers used the attack types DoS and PROBE to target the UDP protocol over the network. They also used the U2R type of attacks to a minor extent that is negligible and R2L attacks were never used. UDP is a very thin layer over IP with less features and complexities compared to TCP. It is affected by five different attack types in the KDD cup training dataset. The hackers exploit the ICMP protocol majorly using the DoS attacks which is the highest of all the four categories of attacks and most affected in all of the protocol types, followed by PROBE. ICMP protocol was not exploited using the U2R and R2L attack types. ICMP is affected by seven different attack types in the KDD cup training dataset. Because of the lack of validation, if attackers want the victim to set its routing information in a particular way, they can send spoofed ICMP redirect messages to the victim, and trick the victim to modify its routing table.

5. Conclusion

Clustering is a powerful data mining technique for data analysis in particular when the data are large and complex. In this paper a key contribution is to make 1000 clusters on 10% of KDD cup Training Dataset. This is fact that data mining gives hidden patterns and explores the data in a different manner. As our dataset is huge (494,020 records) clustering gives more significant advantage. We used ODM tool to build 1000 clusters which worked fine in accomplishing the task. K-means clustering algorithm using ODM has not yet been applied for analyzing the KDD Cup 99 Dataset. We use this effective technique to analyze and study the pattern of attack types in relation with the protocols.

After clustering the training dataset we found some interesting facts. We try to build the relation between attack over the network and protocols used by the hacker. The TCP protocol is more prone to attacks i.e., almost 19 out of the 22 attacks (including DoS, U2R, R2L and PROBE) were used by the hackers. Out of all the attacks made on the TCP protocol 51% were DoS attacks followed by PROBE (28%), R2L (11%) and U2R(8%). Of all the attacks made on ICMP 62% were DoS attacks and 38% were PROBE attacks, U2R and R2L attacks were not used by the hackers. UDP was most affected by PROBE attacks (60%) followed by DoS (35%) and U2R (5%). DoS and PROBE attacks were most prevalent on all the three protocols.

References

[1] H. Oh, I. Doh and K. Chae, “Attack classification based on data mining technique and its application for reliable medical sensor communication”, International Journal of Computer Science and Applications, vol. 6, no. 3, **(2009)**, pp. 20-32.

[2] N. Ye and X. Li, “A Scalable Clustering Technique for Intrusion Signature Recognition”, Proceedings of 2001 IEEE Workshop on Information Assurance and Security, **(2001)**.

[3] G. Singh, F. Masseglia, C. Fiot, A. Marascu and P. Poncelet, “Data Mining for Intrusion Detection: from Outliers to True Intrusions”, The 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD’09), Thailand, **(2009)**.

[4] K. Faraoun and A. Boukelif, “Genetic Programming Approach for Multi-Category Pattern Classification Applied to Network Intrusions Detection”, The International Arab Journal of Information Technology, vol. 4, no. 3, **(2007)**.

[5] W. Lee, S. Stolfo, P. Chan, E. Eskin, W. Fan, M. Miller, S. Hershkop and J. Zhang, “Real Time Data Mining-based Intrusion Detection”, Proceedings of DISCEX II, **(2001)** June.

[6] K. Ioannis, T. Dimitriou and F. C. Freiling, “Towards Intrusion Detection in Wireless Sensor Networks”, 13th European Wireless Conference, Paris, **(2007)** April.

[7] D. Farid, J. Darmont, N. Harbi, N. Hoa and M. Rahman, “Adaptive Network Intrusion Detection Learning: Attribute Selection and Classification”, International Conference on Computer Systems Engineering (ICCSE 09), Bangkok, Thailand, **(2009)** December.

[8] J. Zhang and M. Zulkernine, “Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection”, Symposium on Network Security and Information Assuranc-Proc. of the IEEE International Conference on Communications (ICC), Istanbul, Turkey, **(2006)** June.

[9] M. Tavallaee, E. Bagheri, W. Lu and A. Ghorbani, “A Detailed Analysis of the KDD’99 CUP Data Set”, The 2nd IEEE Symposium on Computational Intelligence Conference for Security and Defense Applications (CISDA), **(2009)**.

[10] M. Campos and B. Milenova, “Creation and Deployment of Data Mining-Based Intrusion Detection Systems in Oracle Database 10g”, an online document at http://www.oracle.com/technology/products/bi/odm/pdf/odm_based_intrusion_detection_paper_1205.pdf.

[11] K. Prothives and S. Srinoy, “Integrating ART and Rough Set Approach for Computer Security”, Proceedings of the International Multi Conference of Engineers and Computer Scientists, vol. 1, **(2009)**.

[12] H. Güneş Kayacık, A. Nur Zincir-Heywood and M. I. Heywood, “Selecting features for intrusion detection: a feature relevance analysis on KDD’99 intrusion detection datasets”, Third Annual Conference on Privacy, Security and Trust, **(2005)** October.

[13] M. Amini and R. Jalili, “Network-based intrusion detection using unsupervised adaptive resonance theory (ART)”, Proceedings of the fourth conference on engineering of intelligent systems (EIS 2004), Madeira, Portugal, **(2004)**.

[14] J. Xiao and H. Song, “A Novel Intrusion Detection Method Based on Adaptive Resonance Theory and Principal Component Analysis”, Proceedings of the 2009 International Conference on Communications and Mobile Computing, vol. 3, **(2009)**.

[15] KDD'99 Competition Dataset, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, **(1999)**.

[16] S. Kumar, "Smurf-based Distributed Denial ofService (DDoS), Attack Amplification in Internet", Second 76 International Journal of Computer Sci ence and Technology.

[17] T. Soni Madhulatha, "An Overview on Clustering Methods", IOSR Journal of Engineering, vol. 2, no. 4, **(2012)** April, pp. 719-725.

[18] J. F. Nieves, "Data Clustering for Anomaly Detection in Network Intrusion Detection", Research Alliance in Math and Science. http://info.ornl.gov/sites/rams09/j_nieves_rodrigues/Documents/report.pdf, **(2009)** August 14.

[19] K. Labib and V. Rao Vemuri, "Detecting Denial-of-Service And Network Probe Attacks Using Principal Component Analysis".

[20] E. Skoudis, "Counter Hack: A Step-by-Step Guide to Computer Attacks and Effective Defenses", Prentice Hall Inc., **(2002)**.

[21] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, vol. 1, **(1967)**, pp. 281-297.

[22] Oracle Technology Network tool. Obtained through internet: http://www.oracle.com/technology/products/bi/odm/odminer.html. Accessed on 15-8-2012.

[23] T. Eldos, M. Khubeb Siddiqui and A. Kanan "On the KDD'99 Dataset: Statistical Analysis for Feature Selection", Journal of Data Mining and Knowledge Discovery, **(2012)**.