International Journal of Information, Communication and Computing Technology
doi:10.5958/2347-7202.2017.00005.6 Jagan Institute of Management Studies, New Delhi

Clustering Algorithm of Data Mining to Detect Network Statistics

Kirti Jain

Jagannath University NCR Haryana India

ABSTRACT

The augmentation growth in network data is causing a contemplative problem of detecting the useful information from the network. In reality the security technologies are not the final solution to prevent from security breaches. The important role of data mining is to detect the patterns of attacks in the network. Since, with a lot of technological preferment, the distinct numbers of attacks are proliferating day by day. Now, cryptography is not commensurate to save the supersensitive information. On the ground to interdict the network attacks their skeletons are identified from KDD 1999 dataset scenting the sequestered data of user's interest in a very less execution time using the data mining tool WEKA. In this proposed study, we are engaging the distinct clustering algorithms such as Farthest-first, Make-Density, Simple K-Means and Filtered Cluster, procuring the statistics of number of attacks found and what are the percentages of attacks on different network layer protocols.

KEYWORDS

Clustering, farthest first, simple K-means, filtered cluster, farthest-first, make density; attacks;.

1 INTRODUCTION

The agendum of Data mining is to discover patterns and modeling queries, which is habitually concealed the important details by using some reasoning techniques such as pattern matching. In this scenario Cyber security is the band-aid with cyber terrorism for distinct cyber-attacks. Clustering techniques is used to associate distinct cyber attacks and reuse the clusters for encountering the attacks when it occurs. The prognosis of attack skeleton can be used to wrap up implied forthcoming attacks depending on the way through which information learnt about terrorists with the help of email and phone conversations. Also, there exist many threats for non real-time data mining such as for network intrusions .To identify these threats we require real-time data mining. Many research people are detecting the use of data mining in the field of intrusion detection. To investigate these threats there is the need of real-time data mining, which means the results should be generated in real-time, after that models should be built models in real-time. For example, the fraud detection in credit card is a form of real-time data processing. Data mining can be used to analyze web logs and audit trails. On the basis of the results generated by the data mining tool, we can determine all the unauthorized intrusions occurred and all the unauthorized queries have been posed.

To perform efficient research Data mining can be used to identify the data of user interest which predicts the attack skeleton for future use. Knowledge discovery or data mining has the attention of IT industry and society to determine the patterns of network attacks. It also analyze the important information from huge volumes of data which are noisy, fuzzy and dynamic.

In this paper clustering technique of data mining is used to detect the attack rule skeleton. In cluster analysis, objects of data are clustered together on the basis of data and relationship between them. There are various algorithms of clustering such as simple K-Means, Farthest–First, Filtered cluster, Make density algorithms of clustering are used different attack structure are represented by using KDD 1999 dataset. The performance of different algorithms is compared and clustered instances formed by distinct algorithms are recorded. The architecture of proposed system is to find the solution of above problems.

In the section II the related work is explained. In section III the multitudinal clustering techniques are described. In section IV the multitudinal algorithms studied are explained how they are useful in detecting intrusion over the networks. In section V the flow diagram of proposed system along with algorithm is explained. Finally, section VI gives the conclusion of the proposed system.

2. RELATED WORK

The apprehension of Intrusion detection was imported by **James Anderson** in 1980 elucidated an intrusion attempt or threat to be potential possibility of a deliberate unauthorized attempt to entree instruction, or relinquish a erroneous system or inutile. Sights hauled for adopting data mining in appeased of NIDS in the late of 1990's. Researchers swiftly perceived the devoir for perseverance of consistent dataset to train IDS tool. Minnesota Intrusion Detection System (MINDS) mingles signature based tool with data mining

[1] *Assistant Professor, Department of Computer Science & IT, Jagannath University, Delhi NCR*
Email:kirti27jain@gmail.com

tactics. Signature based tool (Snort) is worn for misuse detection & data mining for anomaly detection. [1]

Jake Ryan et al adapted neural networks to encounter intrusions. Neural network can be worn to learn a print (user behavior) & diagnose exclusive user. If it does not bout then the system administrator can be forewarned. A back propagation neural network called NNID was competent for this evolution.

Denning D.E et al has devilled a model for overseeing audit trace for abnormal actions in the system. Succulent guidelines are worn to seizing the behavior of user [8] over time. To store the patterns for activities of user deviates significantly from the specified rules, a Rule based system is required. The system which was achieved is of higher precision to identify the record type whether it is normal or attack.

Dewan M et al represents the classification on alert to reduce the false positives in IDS with the help of remodeled self adaptive Bayesian algorithm (ISABA). It is correlated with the security dominion of network intrusion detection based on anomaly.

Aly Ei-Senary et al has integrate the Kuok's algorithms & Apriori data mining algorithm in order to generate the fuzzy logic rules which captures the characteristics of network traffic

3. PROPOSED METHODOLOGY

The General steps in this proposed system of clustering are as follows: In this research work the 10 percent of KDD 1999 dataset for analysis of distinct attack patterns and time taken to perform the distinct clustering process for the attack patterns is recorded. In order to perform clustering first we divide the10 percent of KDD 1999 dataset into 16 equal parts then preprocess the data, perform distinct clustering algorithms such as farthest first, filtered cluster, simple K-Means and density based cluster Algorithm.

The algorithm for this process is explained below along with architecture in figure 1.

Step 1 Convert the dataset into .xls and .csv format for WEKA tool.

Step 2 Divide the dataset D= {d1, d2, d3,d4…….d16} contains 493055 instances.

Step 3 Preprocess the data using the WEKA tool of data mining such that 467224 instances are obtained.

Step 4 Note down the count of distinct instances of attacks by different protocol.

Step 5 Perform the different clustering algorithms and note down the time taken by each algorithm shown in table 2.

Step 6 Then it will result into distinct clusters C= (c1, c2).

Step 7 Detect the patterns obtained through clustering algorithms as shown in Table 1.

Step 8 Classify the different attacks patterns into DOS, Probe, R2L, U2R on the basis of their rule structure.

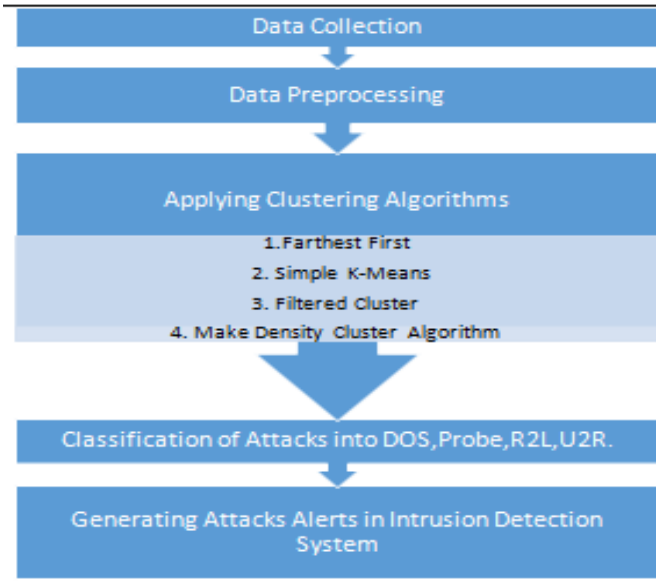

Figure1. Showing the process of algorithm.

4. CLUSTERING TECHNIQUES

Cluster Dissection is one of the most important data mining automation to dislocate the objects of data into disparate allusive subclasses, such that the components belonging to similar clusters are considerably correlative and components belongs to disparate clusters are perfectly contradistinctive to one another [7] Consequently this approach is enforced for segregating log data and revealing unauthorized impressions. As Clustering Technique is called an unsupervised learning approach of data mining this yields an unlabeled data points and whack to associate those data points according to their proximity.

4.1 K-MEANS ALGORITHM

K-means methodology dislocates N vectors into K classes. It is a dynamic clustering process usually starts with an antecedent segregation then use an iterative control process for the optimization of objective function. K-means depicts the cluster condition whether it is normal or under any attack as the normal data is in greater number while constructing

the preliminary dataset. So ,the cluster whose record count is greater than thresholds and it is noted as normal for every TCP/IP connection there are 42 distinct qualitative and quantitative features are extricated.. Some characters are basic (e.g.: protocol type, duration etc), and other characters are accomplished through domain knowledge (e.g.: login attempts which are not successful etc).Every details consists of 41 characteristic attribute, out of which 8 are of string type and remaining are of number type [16]. As network layer has TCP, UDP and ICMP protocols the figure 4.1 shown the clusters of attacks found on these protocol implementing K-means methodology of clustering.

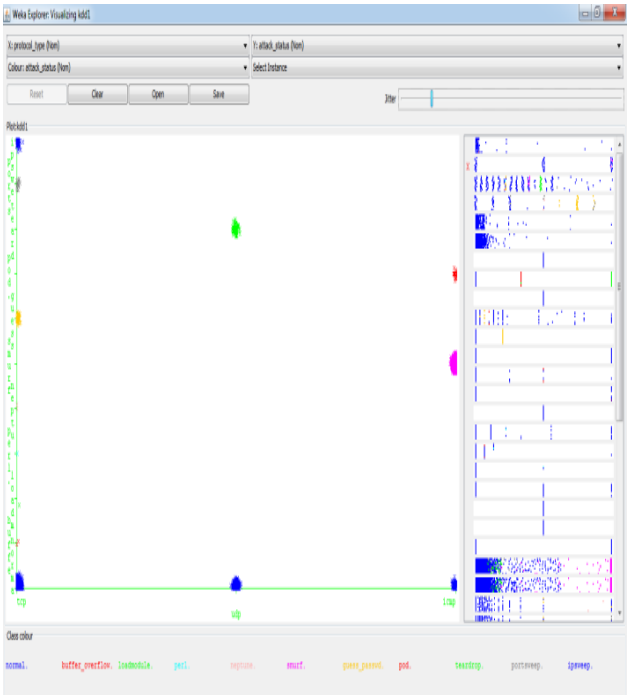

Figure 4.1 K-Means clustering methodology indicates 10 clustered instances of distinct network attacks shown with different colors.

The data fragment is as follows:
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1
,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,253,0.99,0.01,0.00
,0.00,0.00,0.00,0.00,0.00, normal.

0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2
,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,254,1.00,0.01,0.00
, 0.00,0.00,0.00,0.00,0.00, snmpgetattack.

In this fragment the meaning of last record is a normal conjunction or intrusion, for intrusion it points the intrusion type. While performing clustering, it is neglected in order to verify the correctness of result. Attacks are classified into four main categories: U2R (User to Root), such as eject attack; PROBING, such as port scanning attack; DOS (Denial of Service), such as ping of death attack; R2U (Remote to User), such as guest attack.

4.2 FARTHEST-FIRST METHODOLOGY

Farthest-First is performed in two steps: centroid selection and cluster assignment. This methodology is the modification of K-Means algorithm which places every cluster centre at the point extreme far from the current cluster centers. Centroid selection eventuates after selecting the arbitrary data point from the initial cluster center, then selecting the next center as the data point according to the distance metric which is far from the initial center. Subsequently, different centers are chosen in a same manner: they should be far from the set of antecedently chosen centers. Once the *k* number of centroids have been tabbed, this methodology deputize all the other data points to the cluster which are delineated from the immediate centroid and discontinues. Likewise K-Means, Farthest-First methodology needs only a one pass to cluster the group of data points. As there is no average attribute references are figure out to refurbish the centroids. This algorithm has geometric centers of clusters as data points whereas in K-Means all the centriods are the actual data points. This notably increases the efficiency of clustering in all the cases as less realignment and conformance is desired [17].The figure 4.2 shown below has clusters of distinct network attacks on TCP,ICMP and UDP protocol of network layer.

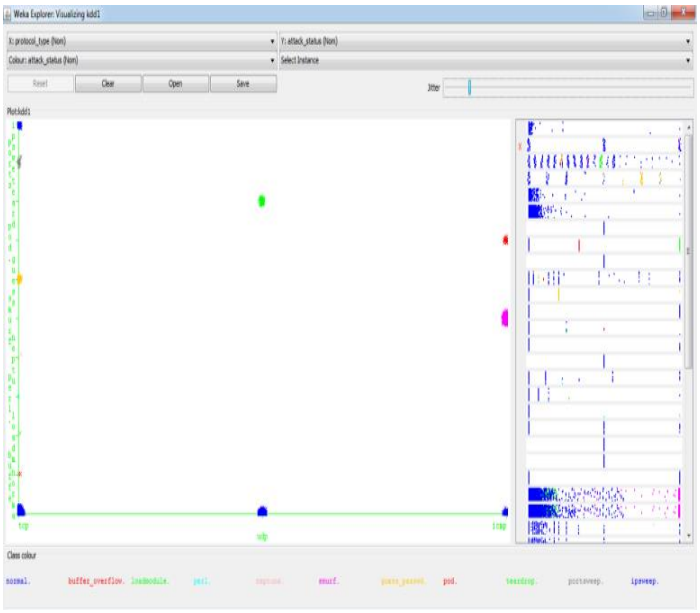

Figure 4.2 Farthest-First methodology indicates the 8 clustered instances of distinct network attacks shown with different colors.

4.3 FILTERED CLUSTER METHODOLOGY

The Filtered Cluster is known to be Meta-Cluster that endeavors the plausibility to implement the undeviatingly filters directly aforetime the Cluster is matriculated. This framework of the filter is solely based upon the preliminary data and trial occurrences are refined by the externally altering their framework. [18]

The algorithm of filtering is as follows:
1. Find the threshold θ of pre-filtering set.
2. Applying clustering to the set of pre-filtered set.
3. Choose the clustering threshold σ, on the basis of the keyword and set the initial relevant document.
4. For every new propaganda α lies within the distance θ from the filtering profile: Increment the propaganda α of the clustering by using the Steps 1 to 3. If the relevant tag for α is found to be true then retrieve that propaganda (α) and correct its relevancy if needed. Filtering of propaganda in the collaboration of viewpoints, data sources and multiple agents is known as collaborative filtering. As shown in figure 4.3 we find the lucid clusters of network attacks employing the Filtered clustering algorithm on network layer.

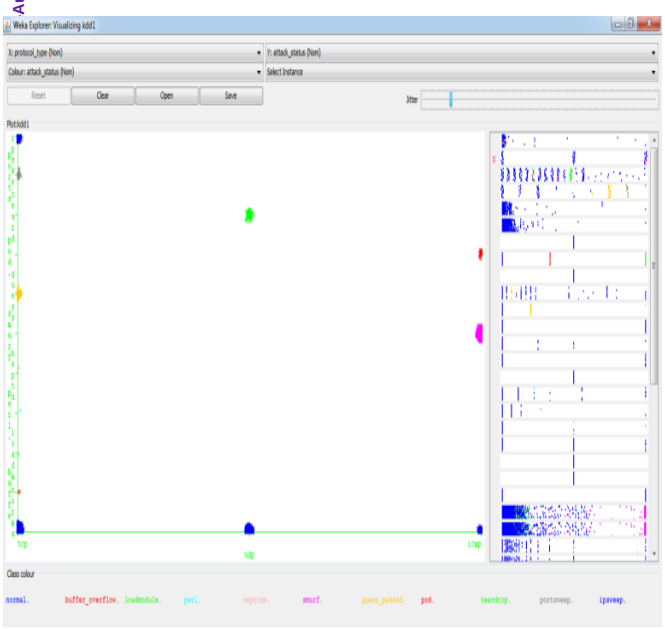

Figure 4.3 Filtered clustering algorithm indicates the 6 clustered instances of network attacks shown with different colors.

4.4 Make density based Methodology

In this methodology, cluster is a condensed shire of points which is spirited by lower condensed domain from the tightly condensed domain. The make density based clustering process can also be worn in noise and when outliers are sustained. The counts with compatible density and present within the comparable region will be coupled to form clusters. Steps are as follows.

1. Enumerate the ε-neighborhood from the data region for all the data objects.
2. Choose CO as core object.
3. Considering every object Ɛ and CO, summate the entire objects from y to CO that are densely united with CO. Repeat until no further y is confronted.
4. Recapitulate step 2 and step 3 as far as all core objects have been refined.

The results of this process is shown in figure 4.4 gives clusters of distinct network attacks on TCP,UDP and ICMP protocol of network layer.

Figure 4.4 Make Density clustering process indicates the 6 clustered instances of network attacks shown with different colors.

5. KDD 1999 DATASET DESCRIPTION

KDD 1999 dataset comprehends a extensive melange of intrusions counterfeit in a military network environment [19] .Each fragment in data is a archive of extracted features from a network connection congregated during the factitious intrusions. A connection is a progression of TCP packets to and from distinct IP addresses. A connection archive comprises of 42 fields. It embodies basic lineaments about TCP connection as duration protocol type, number of bytes transmitted, domain specific features as number of file creation, number of failed login attempts, and whether root shell was obtained. It provides 100,000 labeled data items, composed of 97,276 normal samples and 369948 attack samples. The disparate attributes of this dataset are duration,protocol_type,protocol,sevice,src_bytes,dst_bytes,flag,land,wrong_fragment,urgent,hot,num_failed_logins,logged_in,num_compromised,root_shell,su_attempted,num_root,num_file_creations,num_shells,num_access_files,num_outbo

unds_cmd,is_hot_login,is_guest_login,count,serror_rate,rerror_rate,same_srv_rate,diff_srv_rate,srv_count,srv_serror_rate,srv_rerror_rate,srv_diff_host_rate,dst_host_count,dst_host_srv_count,dst_host_same_srv_rate,dst_host_diff_srv_rate,dst_host_same_src_port_rate,dst_host_srv_diff_host_rate,dst_host_srv_serr,serror_rate,dst_host_rerror_rate,dst_host_srv_rerror_rate,attack_type.

Some of the Characters of dataset are:

1. Basic Characters
It encompasses all the attributes of TCP/IP connection which leads to delay in detection.

2. Characters of Traffic
It is weighing in accordance with window breach & two characters with same host and service.
(a) Characters of Same host
It audits those connections that are from the same destination host.

(b) Characters of same Service
It audits those connections at a particular time breach that posses same service.

3. Content Characters
Probe attack & DOS attack have recurring intrusion persistent impressions than the U2R & R2L. Because these attacks can incorporate many connections to diverse hosts in a distinct time period whereas U2R and R2L percolate only one connection. To expose the above of attacks, dominion expertise is required to access the data belongs to the TCP packets. Ex. Failed login, etc. these Characters are known as content.

The Attacks specimen are classified under distinct attacks which are described as

1. DOS attack – It is a kind of attack where the attacker conceive refining time of the resources and memory busy so as to avoid admissible user from grabbing those resources.

2. U2R attack – Here the attacker sniffs the password or makes some kind of attack to avenue the particular host in a network as admissible user. They can even endorse some susceptibility to yield the root entree of the system.

3. R2L attack – Here the attacker delivers a message to the host in a network over remote system and makes some susceptibility.

4. Probe attack – Attacker will browse the network to congregate information and would make some illegality in the future

6. Experimental Results

The WEKA tool extracts the rule structure of distinct network attacks of KDD 1999 dataset as shown in Table 2 and the percentage of each network attacks found in the dataset is shown in the figure 6.1

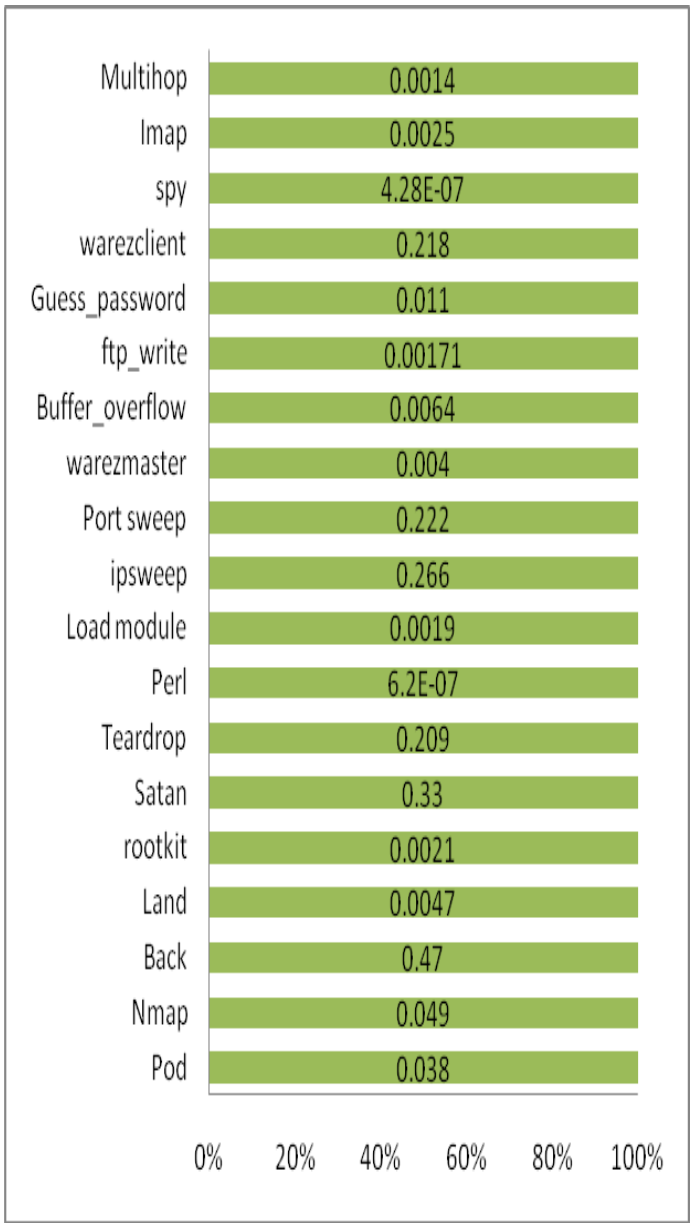

Figure 6.1 Percentage of Attacks found in KDD dataset

Table2. Depicts Rules skeleton of various attacks

S.No	Clustering algorithm	Time taken(in s)	No of instances in cluster 0	No. of instances in cluster 1
1	Farthest First	4.59	408547	84508
2	Filtered Cluster	53.4	267436	225619
3	Simple K-means	53.84	287657	205398
4	Make Density Based Cluster	63.14	288536	204519

Table 1. Showing the clustering Algorithm Performance

S. No	Attack name	Attack Rule skeleton
1	Smurf	ICMP=protocol,ecr_i=service,1032= src_byte,SF=flag, 255= host_count.
2	Neptune	tcp=protocol, private or ctf =service, SO or SF =flag, 1=serror_rate, 1=srv_serror_rate.
3	Phf	tcp protocol,telnet =service, SF =flag, 255=dst_host_count, 0.02=dst_host_serror rate
4	Pod	ICMP=protocol,ecr_i=service,SF=flag,src_byte=1480,1=wrong_fragment,255=,dst_host_count, 0.02=dst_host_diff_srv_rate
5	Nmap	ICMP=protocol,SF or SH =service, 8=rc_byte,1=same_srv_rate, 1=srv_diff_host_rate
6	Back	TCP=protocol, HTTP=service, SF or RSTFR=flag, 54540=src_byte, 7300 or 8314=dst_byte, 1=same_srv_rate, srv_countP5
7	Land	TCP=protocol,Finger=service,SO=flag,1=land,2=srv_count,dst_host_srv_serror_rateP0.17S
8	Rootkit	tcp=protocol, telnet or ftp =service, SF= flag, 255=dst_host_count,0.02=dst_host_diff_srv_rate.
9	Satan	UDP=protocol,private=service,SF=flag,1=src_byte,255=dst_host_count, 1=dst_host_same_src_port_rate.
10	Teardrop	UDP =protocol, SF=service,28=src_byte,3=wrong fragment, 255=dst_host_count.
11	Perl	durationP25,tcp=protocol,telnet=service,SF=flag,1=logged_in,dst_host_srv_count 6 2, dst_host_diff_srv_rate 6 0.07
12	Load module	tcp=protocol,telnet=,service,SF=flag=,1=dst_host_count, 1=dst_host_same_src_port_rate.
13	Ipsweep	icmp =protocol, eco_i =service, SF= flag, 18=src_byte, 1=count, 1=dst_host_count.
14	Portsweep	TCP=protocol, Private or remote_ic =service, 255=dst_host_count,1=dst_host_srv_count
15	Warezmaster	P2duration, tcp =protocol, ftp or ftp_data =service, SF =flag, dst_host_count> 2, dst_host_srv_countP1
16	Buffer_overflow	tcp=protocol,telnet or ftp_data=service, SF=flag, 1=loggin_in,1=dst_host_same_srv_rate.
17	ftp_write	26 or 134=duration,TCP=protocol,FTP or login=service,SF=flag, ,1=logged_in
18	Guess_password	tcp=protocol, telnet=service, RSTO=flag,125 or 126= src_byte, 179=dst_byte,1=hot,1= num_failed_login.
19	Warezclient	tcp=protocol,ftp=service,SF=flag,src_bye>980,dst_byteP1202, hotP3, 255=dst_host_count.
20	Spy	377or299=duration, tcp=service, telnet=flag, 255=dst_host_count, 0.01=dst_host_diff_srv_rate
21	Imap	imap4=service, count 6 4, dst_host_same_srv_rate=1, dst_host_srv_count < =1
22	Multihop	tcp=service,telnet or ftp_data=flag, SF=flag, 63dst_host_srv_count , 1=dst_host_same_src_port_rate

7. CONCLUSION

After analyzing the distinct clustering algorithms and running them under different factors. It is concluded that farthest first algorithm outperform among all the algorithms and 56.31% attacks are from ICMP protocol, 39.47% from TCP and 4.22% from UDP. Thus, we can obtain the following conclusion that the input web dataset consists of 22 attack types that which is further classified into four main domains: Denial of Service (DOS), Probing (Probe), Remote to Local (R2L) and User to Root (U2R).This proposed algorithm is implemented in the future research work. The distinct attacks comes under the four categories of attack is shown in the table 3.

Table 3. Classification of different network attacks

Type	Probe	U2R	DOS	R2L
Attack found in Training Data	Ipsweep, Nmap, Post sweep, Satan	Buffer overflow, Load module, Perl, Root	Pod, Neptune, Back,Land	Imap, Guess_passwd, Ftp_write.

REFERENCES

[1]. G.V. Nadiammai, M. Hemalatha 2014. “Effective approach toward Intrusion Detection System using data mining techniques”, Egyptian Informatics Journal 15, 37–50.

[2]. S.V. Shirbhate, Dr. S.S.Sherekar and Dr. V.M.Thakare 2014 “Performance Evaluation of PCA Filter In Clustered Based Intrusion Detection System”, International Conference on Electronic Systems, Signal Processing and Computing Technologies.

[3]. C.Dartigue, H.I. Jang, and W. Zeng, 2009 “A New Data-Mining Based Approach for Network Intrusion Detection”, Seventh Annual Communications Networks and Services Research Conference, 978-0-7695-3649-1/09 DOI 10.1109/CNSR, IEEE computer Society, PP.372-377.

[4]. S. Revathi, Dr.T.Nalini Adam Prugel-Bennett, Gary Wills 2013. “Performance Comparison of Distinct Clustering Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering 3(2), pp. 67-72.

[5] R. Jingbiao, Y. Shaohong 2010. “Research and Improvement Of Clustering Algorithm in Data Mining”, 2nd International Conference On Signal Processing(ICSPS), 978-1-4244-6893-5,IEEE, PP.VI-842-VI845.

[6] Hongyu Yang, Feng Xie and Yi Lu 2006. “Research on Network anomaly Detection Based on Clustering and Classifier”, IEEE, 1-4244-0605-6/06 pp.592-597.

[7] S. Revathi, Dr.T.Nalini Adam Prugel-Bennett, Gary Wills 2013. “Performance Comparison of Distinct Clustering Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering 3(2), pp. 67-72.

[8] Kingsly Leung Christopher Leckie 2005. “Unsupervised Anomaly Detection in Network Intrusion Detection Clusters”, Conferences in Research and Practice in Information Technology, Vol. 38,PP.333-342.

[9] Pallavi , Sunila Godara 2011. “A Comparative Performance Analysis of Clustering Algorithms”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 1, Issue 3, pp.441-445.

[10] LI Han 2010. “Research and Implementation of an Anomaly Detection Model Based on Clustering Analysis”, International Symposium on Intelligence Information Processing and Trusted Computing.

[11] A.M. Chandrashekhar , K.Raghuveer 2013. “Intrusion Detection Technique using K-Means ,Fuzzy neural networks and SVM classifiers”, International Conference Computer Communication And Informatics(ICCI).

[12]. Narendra Sharma , Aman Bajpai , Mr. Ratnesh Litoriya, 2012.” Comparison the distinct clustering algorithms of Weka Tools International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, Volume 2, Issue.

[13]. E. Hooper 2007. “An Intelligent Intrusion Detection and Response System Using Hybrid Ward Hierarchical Clustering Analysis”, International Conference on Multimedia and Ubiquitous Engineering (MUE'07).

[14] H. Khanbabapour , H. Mirvaziri, 2014. ”An Intelligent Intrusion Detection System Based On Expectation Maximization Algorithm in Wireless Sensor Networks”, International Journal of Information and Communication Technology ,Research Volume 4 No.

[15] S.Shaikh, Ashphak P. Khan , Vinod S. Mahajan 2013. “Implementation of DBSCAN Algorithm for Internet Traffic Classification”, International Journal of Computer Science and Information Technology Research (IJCSITR) Vol. 1, Issue 1, pp: (25-32), October-December.

[16] C. Zhang , G. Zhang , S. Sun, 2009. ” A Mixed Unsupervised Clustering-based Intrusion Detection Model”, Third International Conference on Genetic and Evolutionary Computing.

[17] Pallavi, Sunila Godara 2011. “A Comparative Performance Analysis of Clustering Algorithms”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 1,Issue3, pp.441-445.

[18] Mrutyunjaya Panda1 and Manas Ranjan Patra2, 2009 “A Novel Classification via Clustering Method for Anomaly Based Network Intrusion Detection System”, International Journal of Recent Trends in Engineering, Vol 2, No. 1,PP.1-6.

[19] KDD Cup99 Intrusion Detection Dataset Available from: <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>.