K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset

Vipin Kumar, Himadri Chauhan, Dheeraj Panwar

Abstract— ***Clustering is the most acceptable technique to analyze the raw data. Clustering can help detect intrusions when our training data is unlabeled, as well as for detecting new and unknown types of intrusions. In this paper we are trying to analyze the NSL-KDD dataset using Simple K-Means clustering algorithm. We tried to cluster the dataset into normal and four of the major attack categories i.e. DoS, Probe, R2L, U2R. Experiments are performed in WEKA environment. Results are verified and validated using test dataset. Our main objective is to provide the complete analysis of NSL-KDD intrusion detection dataset.***

Index Terms—Clustering, K-means, NSL-KDD Dataset, WEKA.

I. INTRODUCTION

Human activity nowadays depends on the communication systems. People use communication systems for working, enjoying, business, and governing activities, among others. The majority of the communication systems are interconnected by the global network known as Internet. On Internet every day all kind of users coexist in a public and non-regulated space, and sometimes the activity of a user can cause damage to another user or system. Thus the field of information security has grown and evolved significantly in recent years in order to prevent and to control information security threats. In such case it is necessary to analyze the network elements and network data to determine: the damage or lost caused, the attack method used, the identity of who realized the attack, and the possibility to establish a demand in a court.

Data mining [1] [2] is a helpful practice to uncover new insights, associations and hidden patterns within large data set of logs and messages. Knowledge Discovery in Database (KDD) [3] [4] practice is associated with extraction and discovery of useful information from large relational databases while data mining represents its core as decision support stage. Data mining is the finding process of significant non-intuitive correlations and patterns from a variety of sources, making possible to get high level knowledge from low level data.

Clustering is an unsupervised Data Mining Technique [5-7] where the data set is divided into sub parts sharing common properties. Data Clustering is considered an interesting approach for finding similarities in data and putting similar data into groups [8]. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups.

Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction. By finding similarities in data, one can represent similar data with fewer symbols for example. Also if we can find groups of data, we can build a model of the problem based on those groupings. Another reason for clustering is its descriptive nature which can be used to discover relevant knowledge in huge dataset.

In this paper we present the complete analysis of NSL-KDD Intrusion Detection Dataset using clustering approach. Simple K-Means clustering algorithm is used to identify the different clusters and group them in four major attack categories. We also provide a complete analysis of different kind of attacks present in training and testing dataset. Testing dataset is used to validate the results. Performance of the algorithm is investigated during different execution of the program on the input data set. The execution time for the algorithm is also analyzed.

The remaining paper is structured as follows: Section II summarizes the related work of different researchers. Clustering approach and K-Means algorithm are described in section III. Detailed description of the attacks present in NSL-KDD along with results and analysis is shown in Section IV. Finally the conclusion is summarized in section V.

II. RELATED WORK

Current anomaly detection is often associated with high false alarm with moderate accuracy and detection rates when it’s unable to detect all types of attacks correctly. To overcome this problem, Muda et al. [9] proposed a hybrid learning approach through combination of K-Means clustering and Naïve Bayes classification. They cluster all data into the corresponding group before applying a classifier for classification purpose. An experiment is carried out to evaluate the performance of the proposed approach using KDD Cup ’99 dataset. Result show that the proposed approach performed better in term of accuracy, detection rate with reasonable false alarm rate.

H. Om et al. [10] proposed a hybrid intrusion detection system that combines k-Means, and two classifiers: K-nearest neighbor and Naïve Bayes for anomaly detection. It consists of selecting features using an entropy based feature selection algorithm which selects the important attributes and removes the irredundant attributes. This system can detect the intrusions and further classify them into four categories: Denial of Service (DoS), U2R (User to Root), R2L (Remote to Local), and probe. The main goal is to reduce the false alarm rate of IDS.

Existing IDS techniques includes high false positive and false negative rate. Nadiammai et al. [11] implemented some of the clustering algorithms like k means, hierarchical and Fuzzy C Means, to analyze the detection rate over KDD CUP 99 dataset and time complexity of these algorithms. Based on

Manuscript received on September, 2013.

Vipin Kumar, Dept. of Computer Science and Engineering, Graphic Era University, Dehradun, India.

Himadri Chauhan, Dept. of Computer Science and Engineering, Graphic Era University, Dehradun, India.

Dheeraj Panwar, Dept. of Computer Science and Engineering, DCMTE, Dehradun, India.

evaluation result, FCM outperforms in terms of both accuracy and computational time.

Y. Qing et al. [12] presented an approach to detect intrusion based on data mining frame work. In the framework, intrusion detection is thought of as clustering. The reduction algorithm is presented to cancel the redundant attribute set and obtain the optimal attribute set to form the input of the FCM. To find the reasonable initial centers easily, the advanced FCM is established, which improves the performance of intrusion detection since the traffic is large and the types of attack are various. In the illustrative example, the number of attributes is reduced greatly and the detection is in a high precision for the attacks of DoS and Probe, a low false positive rate in all types of attacks.

The focus of Haque et al. [13] is mainly on intrusion detection based on data mining. The main part of Intrusion Detection Systems (IDSs) is to produce huge volumes of alarms. The interesting alarms are always mixed with unwanted, non-interesting and duplicate alarms. The aim of data mining is to improve the detection rate and decrease the false alarm rate. So, here we proposed a framework which detect the intrusion and after that, it will show the improvement of k-means clustering algorithm.

Poonam et al. [14] compares the performance of the four algorithms on outlier detection efficiency. The main objective is to detect outliers while simultaneously perform clustering operation.

Denatious et al. [15] presents the survey on data mining techniques applied on intrusion detection systems for the effective identification of both known and unknown patterns of attacks, thereby helping the users to develop secure information systems.

III. CLUSTERING APPROACH TO KNOWLEDGE DISCOVERY

The unlabelled data from the large dataset can be classified in an unsupervised manner using clustering algorithms. Cluster analysis or clustering is the assignment of a set of observation into subsets (called clusters) so that observations in same cluster are similar in some sense. A good clustering algorithm results in high intra cluster similarity and low inter cluster similarity. There are three major types of clustering process according to the way they organize data: Hierarchical, Partitioning and Mixture model methods. If the existing data is clustered according to the property of the data, its character and behavior, then cluster impact is valuable. Several data mining techniques have been applied for intrusion detection.

K-Means Clustering: k-means [16-18] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. K-Mean Clustering is unsupervised data mining technique for intrusion detection. It is easy to implement. Three major drawback of K-mean clustering is: class dominance problem, force assignment problem, and no class problem. It has been observed that single model cannot give better result in terms of recall and precision.

The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function known as squared error function.

```
Function Simple k-means ( )
Initialize k prototypes ($w_1, …, w_k$) such that
$w_j = i_j, j \epsilon \{1, … , k\}, l \epsilon \{1, …. , n\}$
Each cluster $C_j$ is associated with prototype $w_j$
Repeat
    for each input vector $i_l$, where $l \epsilon \{1,… n\}$,
        do
            Assign $i_l$ to the cluster $C_{j*}$ with nearest prototype $w_{j*}$
            (i.e., $| i_l - w_{j*} | \le | i_l - w_j |, j \epsilon \{1,…, k\}$)
    for each cluster $C_j$, where $j \epsilon \{1,…, k\}$, do
        Update the prototype $w_j$ to be the centroid of all
        samples currently in $C_j$, so that $w_j = \sum_{il \epsilon Cj} i_l / | C_j|$
    Compute the error function
        $E = \sum_{j=1}^{k} \sum_{i_l \epsilon c_j} | i_l - w_j |^2$
Until E does not change significantly or cluster membership no longer changes.
```

IV. EXPERIMENTS AND RESULTS

In this section we describe the experiments and results and analyze the same.

A. Setup

Experiments are performed in WEKA [19] environment using 20% of the NSL-KDD dataset on a standalone machine having core-i3 processor with 4 GB of RAM. WEKA heap size has been increased to 2 GB to load and analyze the dataset. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values [20]. Before applying the clustering algorithm to the input dataset all attributes are normalized to the range 0 - 1. Number of cluster is set up to four and epochs are set to 100.

Table I. List of Attributes

Total Attributes		
Duration	su_attempted	same_srv_rate
protocol_type	num_root	diff_srv_rate
service	num_file_creation	srv_diff_host_rate
flag	num_shells	dst_host_count
src_byte	num_access_file	dst_host_srv_count
dst_byte	num_outbound cmds	dst_host_same_srv_rate
land	is_host_login	dst_host_diff_srv_rate
wrong_fragment	is_gust_login	dst_host_same_src_port_rate
urgent	count	dst_host_srv_diff_host_rate
hot	srv_count	dst_host_serror_rate
num_failed_login	serror_rate	dst_host_srv_serro_rate
logged in	srv_serror_rate	dst_host_rerror_rate
num_compromised	rerror_rate	dst_host_srv_rerror_rate
root_shell	srv_rerror_rate	class

Table III. Attacks in Testing Dataset

Testing Dataset (20 %)	Attack- Type (37)
DoS	Back, Land, Neptune, Pod, Smurf, teardrop, Mailbomb, Processtable, Udpstorm, Apache2, Worm,
Probe	Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint
R2L	Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Xlock, xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps

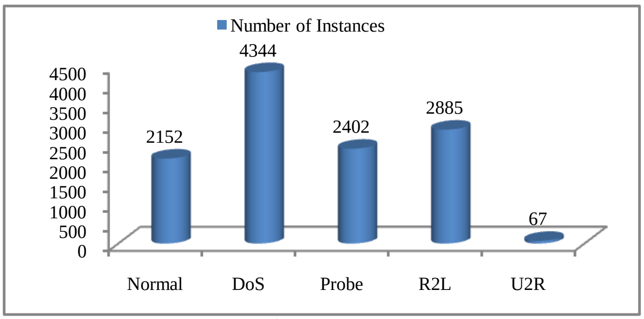

Fig II: Number of Instances in Testing Dataset

B. *Preprocessing Dataset*

The KDD Cup'99 [21] intrusion detection dataset suffers from some of the problems [22] such as redundant records so we conducted experiments on NSL-KDD [23]. NSL-KDD consists of selected records of the complete KDD data set and is publicly available for researchers. In each connection are 41 attributes describing different features of the connection and a label assigned to each either as an attack type or as normal. Table 1 shows all the attributes present in the NSL-KDD dataset.

Table II. Attacks in Training Dataset

Training Dataset (20 %)	Attack – Type (21)
DoS	Back,Land,Neptune,Pod,Smurf,teardrop
Probe	Satan,Ipsweep,Nmap, Portsweep,
R2L	Guess_Password, Ftp_write,Imap, Phf,Multihop, Warezmaster, Warezclient, Spy
U2R	Buffer_overflow, Loadmodule, Rootkit

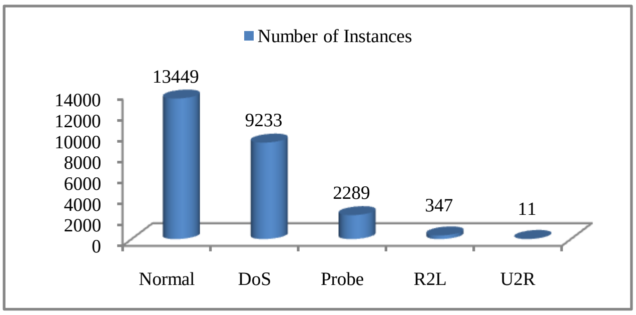

Fig I: Number of Instances in Training Dataset

The training dataset is made up of 21 different attacks out of the 37 present in the test dataset. The known attack types are those present in the training dataset while the novel attacks are the additional attacks in the test dataset i.e. not available in the training datasets. The attack types are grouped into four categories: DoS, Probe, U2R and R2L. The training dataset consisted of 25193 instances among which 13449 are normal and 11744 are attack type whereas the testing dataset consist of 2152 normal and 9698 attack types.

C. Result Analysis

We have analyzed the NSL-KDD dataset using Simple K-means clustering algorithm. We have also presented all the major attack types present in training and testing dataset and are shown in table II and table III. Performance of the K-Means algorithm is evaluated using Euclidean Distance measure. Tables and figure below depicts the outcome of the experiments. Instances are distributed over four clusters: Cluster 1 to Cluster 4 including normal instances. Figure III shows the distribution of the instances in different categories i.e. Normal, DoS, Probe, R2L, U2R. K-means algorithm takes 9.25 seconds to build cluster models and the mean squared error in this process is 19308.72. Clusters shown in figure III give a clear representation of all the instances and their belongingness to their categories. Number of instances in each cluster is shown in table IV. It also represents the distribution of instances to particular attack type. K-means clustering algorithm provides better distribution of instances to their categories as presented in the NSL-KDD dataset. Experimental results are validated using test dataset.

Table IV. Distribution of Instances to Clusters

Cluster	No. of Instances in Each Cluster	%	Normal	Dos	Probe	R2L	U2R
Cluster 1	5129	20.35%	3241	755	1163	9	1
Cluster 2	10025	39.79%	9623	196	8	188	10
Cluster 3	7027	27.89%	35	6901	89	2	0
Cluster 4	2971	11.79%	550	1382	1029	10	0

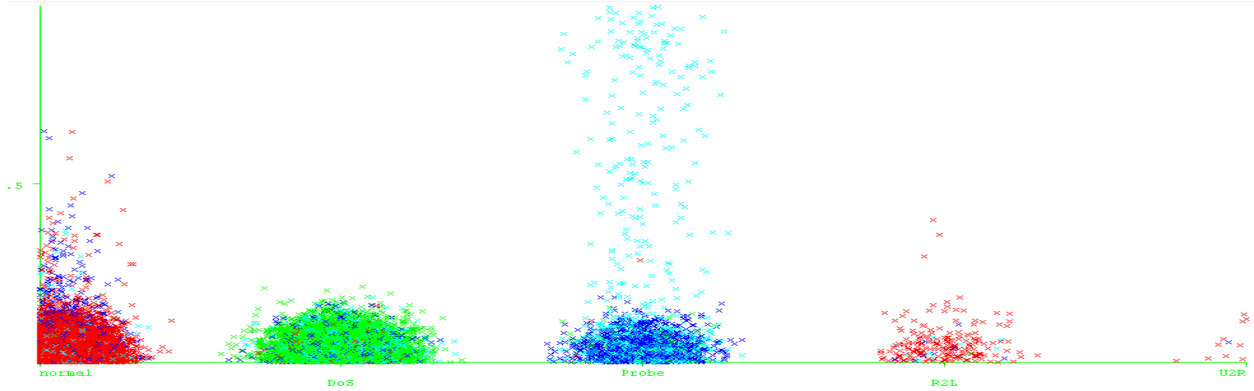

Fig III. Cluster representation of four types of attack with normal traffic.

V. CONCLUSION

In this paper we have analyzed NSL-KDD dataset to using K-means clustering. Clustering algorithms proves to be very useful when we have huge amount of unlabelled dataset. The study analyses the different types of attacks present in NSL-KDD. K-means Clustering applied here is able to efficiently detect new type of attacks present in dataset. K-means clustering is able to cluster the attacks present in training dataset into four major categories giving a better representation of the clusters. The main objective of the paper was to provide a complete analysis of the NSL-KDD dataset and the attacks presented. We used K-means algorithm for this purpose and also represented the distribution of instances in clusters providing better representation of the instances and making it clearer to understand.

In future, an association rule based approach or IF-THEN rules could be effective in categorizing the traffic in different classes. However accuracy of the algorithms plays an important role to correctly cluster the datasets. Standalone algorithms may not be able to provide efficient results. A hybrid approach to data clustering can also be applied for analysis and to obtain low inter-cluster similarity.

REFERENCES

[1] Lee W., and Stolfo S. J., (1998): “Data mining approaches for intrusion detection”, 7th USENIX Security Symposium, San Antonio, 635-660.

[2] Gaol F. L., Yi S., Deng F., (2012): “Research of Network Intrusion-Detection System Based on Data Mining”, Recent Progress in Data Engineering and Internet Technology, vol. 157: 141-148. Springer Berlin, Heidelberg.

[3] Fayyad U. M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R., (1996): editors, “Advances in Knowledge Discovery and Data Mining”, MIT Press.

[4] Kaufman K. A., Michalski R. S. and Kerschberg L., (1991): “Mining for Knowledge in Databases: Goals and General Description of the INLEN System”, in G. Piatetsky-Shapiro and W. J. Frawley (eds.), *Knowledge Discovery in Databases, AAAI/MIT Press*, 449-462.

[5] Duda, R. O., Hart P. E., and Stork D.G., “*Unsupervised Learning and Clustering”*, Chapter 10 in *Pattern classification* (2nd edition), p. 571, New York, NY: Wiley, ISBN 0-471-05669-3.

[6] Fisher D. H., (1987): “Knowledge Acquisition via Incremental Conceptual Clustering”, *Machine Learning* 2: 139-172.

[7] Zanero S., Savaresi S. M., (2004): “Unsupervised learning techniques for an intrusion detection system”, ACM Symposium on Applied Computing (SAC’04), Nicosia, Cyprus.

[8] Jang J. S. R., Sun C. T., and Mizutani E. (1997): "Neuro-Fuzzy and Soft Computing-A Computational Approach to Learning and Machine Intelligence [Book Review]," *Automatic Control, IEEE Transactions on*, vol. 42, no. 10, pp. 1482-1484.

[9] Muda Z., Yassin W., Sulaiman M. N., and Udzir N. I., (2011): "Intrusion detection based on K-Means clustering and Naive Bayes classification," in *Information Technology in Asia (CITA 11), 7th International Conference on*, Kuching, Sarawak, pp. 1-6.

[10] Om H., and Kundu A., (2012): "A hybrid system for reducing the false alarm rate of anomaly intrusion detection system," in *Recent Advances in Information Technology (RAIT), 1st International Conference on*, Dhanbad, pp. 131-136.

[11] Nadiammai G. V., and Hemalatha M., (2012): "An Evaluation of Clustering Technique over Intrusion Detection System," in *International Conference on Advances in Computing, Communications and Informatics (ICACCI'12)* Chennai, pp. 1054-1060.

[12] Ye Q., Wu X., and Huang G., (2010): "An intrusion detection approach based on data mining," in *Future Computer and Communication (ICFCC), 2nd International Conference on*, Wuhan, pp. V1-695-V1-698.

[13] Haque M. J., Magld K. W., and Hundewale N., (2012): "An intelligent approach for Intrusion Detection based on data mining techniques," in *Multimedia Computing and Systems (ICMCS), International Conference on*, Tangier, pp. 12-16.

[14] Poonam P., and Dutta M., (2012): "Performance Analysis of Clustering Methods for Outlier Detection," in *Advanced Computing & Communication Technologies (ACCT), Second International Conference on*, Rohtak, Haryana, pp. 89-95.

[15] Denatious D. K., and John A., (2012): "Survey on data mining techniques to enhance intrusion detection," in *Computer Communication and Informatics (ICCCI), International Conference on*, Coimbatore, pp. 1-5.

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

[17] Velmurugan T., and Santhanam T., (2010): "Performance Evaluation of K-Means and Fuzzy C-Means Clustering Algorithms for Statistical Distributions of Input Data Points," *European Journal of Scientific Research*, vol. 46, no. 3, pp. 320-330.

[18] Borah S., and Ghose M. K., (2009): "Performance analysis of AIM-K-Means and K-Means in quality cluster generation," *Journal of Computing*, vol. 1, no. 1.

[19] WEKA (2008): Data Mining Machine Learning Software [Available Online] http://www.cs.waikato.ac.nz/ml/weka/

[20] Chandrashekhar A. M., and Raghuveer K., (2012): "Performance evaluation of data clustering techniques using KDD Cup-99 Intrusion detection data set," *International Journal of Information & Network Security (IJINS)*, vol. 1, no. 4, pp. 294-305.

[21] “KDDCup 1999 Dataset”. [Available online]: http://kdd.ics.uci.edu/databases/kddcup1999.html/

[22] McHugh J., (2000): “Testing Intrusion detection system: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory”, *ACM Transaction on Information and system security*, vol. 3, no. 4, pp.262-294.

[23] NSL-KDD dataset, (2006): [Available Online] http://iscx.ca/NSL-KDD/