Anomaly-based Intrusion Detection System Using K-Means Clustering

Mohamed Sylla,
Joana EWUNTOMAH, Mouhamadou De Gaulle BODIANG
African Institute For Mathematical Sciences
(AIMS-Sngal 2018-2019)

***Abstract*—The need to secure networks has increased as the number of people connecting to the network is growing rapidly and uses networks to store or access critical information. In this article, we evaluated the machine learning algorithm that is the K-means clustering and compared our results to the existing results, this clustering algorithm is efficient and allows detection 'intrusion'. This algorithm allows the reduction of false alarms. The data set used in this project would be the database containing the KDD '99 data set to be evaluated, which includes a wide variety of intrusions**

I. Introduction

THE amount of data stored on personal computers, organizational and government computers is increasing more and more as the days goes by. This calls for need for individuals to be extra careful in order not to lose valuable data or having their computers flooded with data to jam the network traffic. Valuable information are always attractive to attackers, they are therefore on alert to attack networks systems which requires intelligent anomaly detection systems to defend the network system. Two approaches exist in the literature and in existing tools, signature detection and anomaly detection (Rafath Vasumathi, 2017) in [1]. Signature-based systems are very effective at detecting network breaches of known signatures. However, they cannot defend networks against unknown anomalies. Intrusion is when an attacker enters a system with the aim of stealing or modifying important and confidential information. This occurs when malicious packets are sent to the network system.

There is a system to detect intrusions in a network called Intrusion detection system (IDS) ahead popularized by James P. Anderson in 1980 (Anderson, 1980) in [2]. Traditional Signature based automatic detection methods have been widely used in intrusion detection system. They suffer from long response time and inability to detect new unknown attacks. Anomaly intrusion detection system is used for detecting both network and computer intrusions by classifying them as either normal or anomalous. Anomaly detection uses non-anomalous traffic to build a normal traffic pattern. The main aim of this research is to improve anomaly intrusion detection system using K-means clustering to give a better rate of anomaly detection. This helped reduced false alarm rates and increase false positive rate. This paper is divided in five sections starting from introduction, followed by related work, our methodology in details with the experimental results and conclusion.

II. KDD CUP 99 Data Set

The KDD training data set consists of 10% of the original dataset, or about 494,020 simple connection vectors, each containing 41 entities that label exactly one type of specific attack, ie normal or attack. Each vector is called normal or attack, with exactly one type of specific attack. Deviations from "normal behavior", anything that is not "normal", are considered attacks. [3] Normal attacks are records with normal behavior. The training data set contains 19.69% of normal attack connections and 80.31%. KDD CUP 99 was the most widely used in attacks on the network. The simulated attack falls into one of the following four categories [4]: **1. Denial of service attack (DOS):** the attacker generates computing resources or memory too busy or too saturated to treat a legitimate request, or to refuse the access of the legitimate users to the machine. DOS includes attacks: 'neptune', 'back', 'smurf', 'pod', 'land', and 'tear'.

2. Root Attack Users (U2R): The attacker begins by accessing a normal user account on the system and is able to exploit a vulnerability to gain root access to the system. U2R includes attacks: 'buffer_overflow', 'loadmodule', 'rootkit' and 'perl'

3. Remote to Local Attack (R2L): The attacker sends packets to the machine over a network, but does not have an account on that machine and exploits a vulnerability to gain local access as an attacker.

user of this machine. R2L contains the attacks: warezclient, multihop, ftp_write, imap, guess_passwd, warezmaster, spy, and phf

4. Probing Attack (PROBE): The attacker attempts to collect information about the computer network for the apparent purpose of circumventing its security. PROBE includes attacks: 'portsweep', 'satan', 'nmap' and 'ipsweep'

The main objectives of network intrusion detection are the recognition of rare attack types such as U2R and R2L, increased detection rate of accuracy of suspicious activity and improved efficiency of detection models. intrusion in real time. This finds that the training database included 494,019 records, of which 97,277 (19.69%) were "normal", 391,458 (79.24%) were DOS, 4,107 (0.83%) probes, 1 126 (0.23%) R2L and 52 (0.01%) U2R. attacks. Each record has 41 attributes describing different characteristics and a label assigned to each of them,

as an ”attack” or ”normal” type. TCP, UDP and ICMP are the protocols taken into account. They are detailed below:

TCP: TCP stands for Transmission Control Protocol. It is an important protocol for the Internet protocol suite at the transport layer, which is the fourth layer of the OSI model. TCP is a connection-oriented reliable protocol that implies that data sent by a side are sure to reach the destination in the same order. TCP sends data in labeled packets to the network after dividing them. The most common protocols that use TCP are HTTP, SMTP / POP3 / IMAP (messaging) and, FTP. UDP: UDP stands for ”User Datagram Protocol”. Its behavior is similar to that of TCP, except that it is an unreliable and connectionless protocol. Because data travels over unreliable media, data may not arrive in the same order, packets may be missing, and packet duplication may be possible. UDP is a transactional protocol that is useful in situations where the delivery of data within a certain time is more important than the loss of a few packets on the network.

ICMP: ICMP stands for ”Internet Control Message Protocol”. It is basically used for communication between two connected computers. Its main purpose is to send messages on networked computers. ICMP redirects messages and is used by routers to provide up-to-date routing information to hosts, which initially have minimal routing information. The host modifies its routing table based on the ICMP redirect message that it receives.

Various researchers have analysed the KDD Cup 99 Dataset using various methods, Mohammad Khubeb Siddiqui and Shams Naahid[5] applies K-means clustering algorithm using Oracle Data Miner(ODM).

III. RELATED WORKS

This section provides a detailed study on classification based anomaly detection methods and related application domains.Clustering based Anomaly Detection Techniques **Gerhard Mnz, Sa Li et al** in [6] have proposed a novel flow-based anomaly detection scheme based on the K-mean clustering algorithm Training data containing unlabeled flow records are separated into clusters of normal and anomalous traffic. The corresponding cluster centroids are used as patterns for computationally efficient distance-based detection of anomalies in new monitoring data on KDD 99 data se. And they give an introduction to Network Data Mining, i.e. the application of data mining methods to packet and flow data captured in a network and the anomaly detection processes. **Z. Muda et al** in [7] have proposed an hybrid learning approach through combination of K-Means clustering and Nave Bayes classification. The proposed approach will be cluster all data into the corresponding group before applying a classifier for classification purpose. they using the data set KDD Cup 99 for evaluate the performance.

Yi Yi Aung * , Myat Myat Min in [8] give highlights the similar distribution of attacks nature by using K-means and also the effective accuracy of Random Forest algorithm in detecting intrusions. They describes full pattern recognition and machine learning algorithm performance for the four attack categories, such as Denial-of-Service (DoS) attacks (deny legitimate request to a system), Probing attacks (information gathering attacks), user-to-root (U2R) attacks (unauthorized access to local super-user), and remote-to-local (R2L) attacks (unauthorized local access from a remote machine) shown in the KDD 99 Cup intrusion detection dataset.

Vipin Kumar, Himadri Chauhan, Dheeraj Panwar in [9] have proposed trying to analyze the NSL-KDD dataset using Simple K-Means clustering algorithm. they tried to cluster the dataset into normal and four of the major attack categories i.e. DoS, Probe, R2L, U2R. they experiments are performed in WEKA environment. so they give provide the complete analysis of NSL-KDD.

Amuthan et al in [10] The k-Means algorithm groups N data points into k disjoint clusters, where k is a predefined parameter. Network intrusion detection system aims to identify attacks or malicious activity in a network with a high detection rate while maintaining a low false alarm rate. In this paper the autor , K-Means + C4.5, detection method was a method used to cascade k-Means clustering and the C4.5 decision tree methods for classifying anomalous and normal activities in a computer network using the divide-and-conquer algorithm. KDD99 data set for conducting the experiments
(Mohammad Khubeb Siddiqui and Shams Naahid) [5]In this paper the author make an analysis of 10 of KDD cup99 training dataset based on intrusion detection. He 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; they used the Oracle 10g data miner as a tool for the analysis of dataset and build 1000 clusters to segment the 494,020 records.

IV. METHODOLOGIE

This section includes the data mining classification algorithm used, namely K-means.
-K-medium clustering algorithm K-means Clustering Algorithm Clustering, based on distance measurements made on objects and classify objects (invasions) into clusters. In its excursion, the algorithm has no information about the label of the learning data.
- We need a notion of similarity or distance
- Do we need to know a priory how many clusters exist?

Measurement of distance or similarity plays an important role in collecting observations into homogeneous groups. Jacquard affinity measurement, the longest common order scale (LCS), is important that the event is to awaken the size to determine if normal or abnormal The Euclidean distance is Euclidean dimensions, distance size widely used for vector space. The Euclidean distance can be defined as the square root of the total difference of the same vector dimension.

In this article, we use the K-means algorithm to group data set connections. The K-means algorithm is one of the most widely recognized classification tools. K-means groups data

according to their characteristic values into a specified number of distinct K clusters. Data classified in the same cluster has identical functionality values. K, the positive integer indicating the number of clusters, must be provided in advance. The steps involved in a K-averaging algorithm are given accordingly:

1. The K points indicating the data to be grouped are placed in the space. These points designate the centroids of the primary group.
2. Data is assigned to the group adjacent to the center of gravity.
3. The positions of all K centroids are recalculated as soon as all data are assigned.
4. Repeat steps 2 and 3 until the center of gravity remains unchanged. This causes the data to be partitioned into groups.

The partition of the pre-processed dataset is done using K-means algorithm with the K value as 10. Because we have the dataset that contains normal attack categories and 4 such as DoS, Probe, U2R, R2L.

V. EXPERIMENTS AND RESULTS

A. INFORMATION DATASET

KDD dataset [11] covered four major categories of attacks which is Probe, DoS, R2L and U2R. In order to demonstrate the abilities to detect different kinds of intrusions, the training and testing data covered all classes of intrusion categories as listed in the following as adopted from the [16].
Table I summarizes the distribution records for training dataset according to class type.And testing dataset is also used.

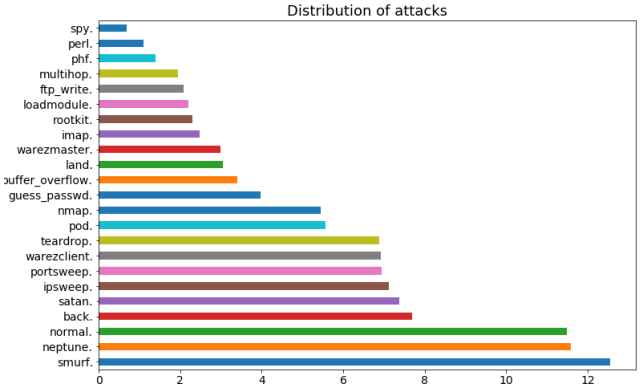
TABLE I. SAMPLE DISTRIBUTION OF THE TRAINING DATASET AND TEST DATASET

Class	Training Set	percentage	Test Set	Percentage
Normal	97278	19.69	60593	19.4
Probe	4107	0.83	4166	1.33
Dos	391458	79.24%	231455	74.4%
U2l	1126	0.01%	88	0.028%
U2r	52	0.23%	14727	4,74%
Total	494021	100%	311029	100%

In total, 42 features have been used in KDD99 dataset and each connection can be categorized into five main classes (one normal class and four main intrusion classes: probe, DOS, U2R, R2L).

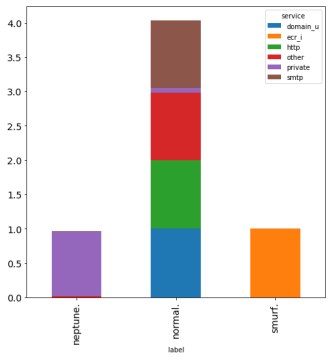
There are 22 different types of attacks that are grouped into the four main types of attacks (probe, DOS, U2R, R2L) tabulated in Table II

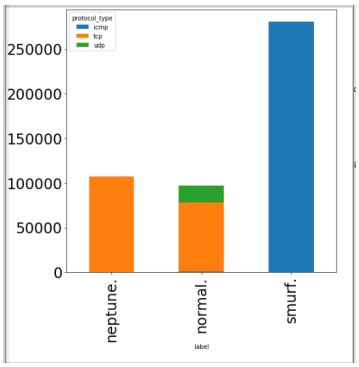
Table II



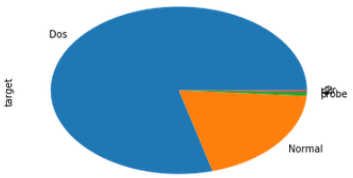
4 Attack Classes	22 Attacks Classes
Probing	ipsweep, nmap, portsweep, satan
Denial of Service (DOS)	back, land, neptune, pod, smurt, teardrop
User to Root (U2R)	buffer_overflow, perl, loadmodule, rootkit
Remote to User (R2L)	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster

We can show the different attacks using the network service. To facilitate viewing, we will only consider attacks that represent at least 1% of the total percentage of attacks or types of network services. and protocol using.





For finish we can see the distribution for all class attacks



B-Evaluation Measurement

An Intrusion Detection System (IDS) requires high accuracy and detection rate as well as low false alarm rate. In general, the performance of IDS is evaluated in term of accuracy, detection rate, and false alarm rate as in the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Detection\ Rate = \frac{TP}{TP + FP} \quad (2)$$

$$FalseAlarm = \frac{FP}{TN + FP} \quad (3)$$

Table III shows the categories of data behavior in intrusion detection category classes (Normal and Attacks)

Actual	Predicted Normal	Predicted Attack
Normal	TN	FP
Intrusions (attacks)	FN	TP

True positive (TP) when attack data detected as attack True negative (TN) when normal data detected as normal False positive (FP) when normal data detected as attack False negative (FN) when attack data detected as normal

C. Result and Discussion

TABLE IV. DETECTION RESULT FOR THE NORMAL AND ATTACK CLASSES USING TESTING DATASET

Actual	Predicted Normal	Predicted Attack
Normal	22750	2292
Intrusions (attacks)	31	6030

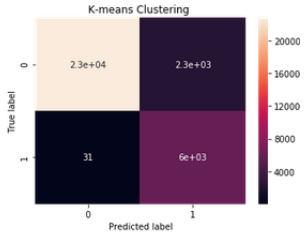


TABLE V. Measur For Evaluation THE NORMAL AND ATTACK CLASSES USING TESTING DATASET

precision	recall	f1-score	support	total
Normal	0.72	0.99	0.84	6061
Intrusions (attacks)	1.00	0.91	0.95	25042
total	0.95	0.93	0.93	31103

Table VIIII shows the measurement in terms of accuracy, detection rate, and false alarm using the training and testing sets of K-Means Clustering. the K-Means Clustering Algorithm recorded high accuracy and detection rate with low false alarm percentage. The clustering techniques used as a pre- classification component for grouping similar data into respective classes helped to produce better results as compared to single classifier. The K-Means Clustering also allows by group data to be classified , hence improving the accuracy and detection rate with acceptable false alarm. For instance, the K-means Clustering enhances the accuracy for alternatively the values k (variation).

TABLE VII. K-MEANS CLUSTERING USING TESTING DATASET

DataSet	Training
Method	K-Means
Accuracy	93
Detection Rate	0,9949
False Alarm	0.0014

Table X show further comparisons made for the proposed another approach using the same KDD Cup 99 dataset as in previous researches in term of accuracy (AC), detection rate (DR), false positive (FP) and false alarm (FA).

TABLE VIII. COMPARISON WITH PREVIOUS FINDINGS USING TESTING DATASET

METHODS	**AC**	**DR**	**FP**	**FA**
K-Means (for k=60)	**93**	**99.49**	**0.0014**	**0.2754**
NBC[19]	88.2	85.0	NA	33.7
RA KM[20]	NA	81.75	NA	26
KM-KNN [17][20]	93.55	98.68	0.98	4.79
HC and SVM [18]	95.7	NA	0.7	NA

Training Time (TT).

TABLE IX. Performance Comparison WITH PREVIOUS FINDING [12]

Classifier	Performance	(TT) Sec.
K-Means	78.7	70.7
K-Means Clustering	**93**	**4.57**
NEA	92.22	10.63
FCC	89.2	56.2
ID3	72.22	120
J48	92.06	15.85
PART	45.67	169
NBTree	92.28	25.88
SVM	81.38	222.28
Fuzzy logic	94.8	873.9
nave Bayes	78.32	5.57
BayesNet	90.62	6.28
Decision Table	91.66	66.24
Random Forest Classifier	92.81	491
JRip	92.30	207.47
OneR	89.31	3.75
MLP	92.03	350.15
SOM	91.65	192.16
GAU	69.9	177.4
MARS	96.5	67.90
Apriori	87.5	18

WE can see we get the good performance that some algorithm and we get 93% in [22].

VI. Conclusion

In this paper we presented an approach for visualizing network attacks data using clustering. It is an easy, simple and fast way of analyzing the flow data. By the help of clustering we can predict the type of flow i.e. attacks or normal by performing some clustering on the particular attributes. We present the K-means Clustering algorithm for anomaly based intrusion detection and apply it by using Cluster Results on a subset of KDD-99 dataset [] showed accuracy of the algorithm. We have done compare some algorithme and With the K-Mean Clustering we have get the objectif it say reduced false alarm rates and increase false positive rate(it is ok)

References

[1] Rafath, S., Vasumathi, D., (2017). Review on anomaly based network intrusion detection system. India: IEEE

[2] Anderson, J.P., (1980). Computer security threat monitoring and surveillance. USA: Technical Report, James P. Anderson Co

[3] 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.

[4] M. Tavallaee, E. Bagheri, W. Lu and A. Ghorbani, A Detailed Analysis of the KDD99 CUP Data Set, The 2nd IEEE Symposium on Computational Intelligence Conference for Security and Defense Applications (CISDA), (2009).

[5] Mohammad Khubeb Siddiqui and Shams Naahid Analysis of KDD CUP 99 Dataset using Clustering based Data Mining, College of Computer Engineering and Sciences, Salman bin Abdulaziz University, Kingdom Saudi Arabia, International Journal of Database Theory and Application-Vol.6, No.5 (2013), pp.23-34

[6] Gerhard Mnz, Sa Li, Georg Carle,Traffic Anomaly Detection Using K-Means Clustering,Computer Networks and Internet Wilhelm Schickard Institute for Computer Science University of Tuebingen, Germany

[7] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir Intrusion Detection based on K-Means Clustering and Nave Bayes Classification, Faculty of Computer Science and Information Technology University Putra Malaysia 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia, 2011 7-th International Conference on IT in Asia (CITA)

[8] Yi Yi Aung * , Myat Myat Min An Analysis of K-means Algorithm Based Network Intrusion Detection System, Faculty of Computer Sciences, University of Computer Studies, Mandalay, UCSM, 0000, Myanmar, Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 496-501(2018)

[9] Vipin Kumar, Himadri Chauhan, Dheeraj Panwar, K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-4, September 2013

[10] Amuthan, P. M., Rajeswarib, R., Rajaramc, R., (2011). Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm. India: Elsevier Ltd. Selection and/or peer-review under responsibility of ICCTSD 2011

[11] KDD. (1999). Available at http://kdd.ics.uci.edu/databases/ - kdd-cup99/kddcup99.html

[12] Safaa O. Al-mamory "Evaluation of Different Data Mining Algorithms with KDD CUP 99 Data Set"University of Babylon/college of computers and Sciences Firas S. Jassim University of Diyla /college of Sciences

[13] M. Tavallaee, E. Bagheri, W. Lu and A. Ghorbani, A Detailed Analysis of the KDD99 CUP Data Set, The 2nd IEEE Symposium on Computational Intelligence Conference for Security and Defense Applications (CISDA), (2009).

[14] 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.

[15] X. Wu, V.Kumar, Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg, Top 10 algorithms in data mining, Survey Paper(2008).

[16] Warusia Yassin, Nur Izura Udzir 1 , Zaiton Muda, and Md. Nasir Sulaiman "ANOMALY-BASED INTRUSION DETECTION THROUGH K-MEANS CLUSTERING AND NAIVES BAYES CLASSIFICATION, "Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia, izura@fsktm.upm.edu.my

[17] C. F. Tsai, and C.Y Lin, A triangle area-based nearest neighbors approach to intrusion detection, Pattern Recognition, 2010, 43(1):222-229.

[18] S-J Horng, M-Y Su and Y-H Chen, A novel intrusion detection system based on hierarchical clustering and support vector machines, Expert Systems with Applications, 2011, 38:306313.

[19] M. Tavallaee, E. Bagheri, W. Lu and A. Ghorbani, A Detailed Analysis of the KDD99 CUP Data Set, The 2nd IEEE Symposium on Computational Intelligence Conference for Security and Defense Applications (CISDA), (2009). KDD. (1999). Available at http://kdd.ics.uci.edu/databases/ - kdd- cup99/kddcup99.html 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] X. Wu, V.Kumar, Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg, Top 10 algorithms in data mining, Survey Paper(2008).

[20] Sonia Nandi, Suman Deb an Mitali Sinha Progress in Advanced Computing and Intelligent Engineering ..., Volume 2 [22]Safaa O. Al-mamory Evaluation of Different Data Mining Algo- rithms with KDD CUP 99 Data SetUniversity of Babylon/college of computers and Sciences Firas S. Jassim University of Diyla /college of Sciences