International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1952332

Implementation of K-Means Clustering for Intrusion Detection

Saba Karim*, Rousanuzzaman, Patel Ayaz Yunus, Patha Hamid Khan, Mohammad Asif

B.E. Computer Jamia Institute of Engineering and Management Studies, Akkalkuwa, Maharashtra India

ABSTRACT

Machine learning is embraced in an extensive variety of areas where it demonstrates its predominance over customary lead based calculations. These strategies are being coordinated in digital recognition frameworks with the objective of supporting or notwithstanding supplanting the principal level of security experts although the total mechanization of identification and examination is a luring objective, the adequacy of machine learning in digital security must be assessed with the due steadiness. With the improvement of the Internet, digital assaults are changing quickly and the digital security circumstance isn't hopeful. Since information are so critical in ML/DL strategies, we portray a portion of the normally utilized system datasets utilized in ML/DL, examine the difficulties of utilizing ML/DL for digital security and give recommendations to look into bearings. Malware has developed over the previous decades including novel engendering vectors, strong versatility methods and different and progressively propelled assault procedures. The most recent manifestation of malware is the infamous bot malware that furnish the aggressor with the capacity to remotely control traded off machines therefore making them a piece of systems of bargained machines otherwise called botnets. Bot malware depend on the Internet for proliferation, speaking with the remote assailant and executing assorted noxious exercises. As system movement, action is one of the principle characteristics of malware and botnet task, activity investigation is frequently observed as one of the key methods for recognizing traded off machines inside the system. We present an examination, routed to security experts, of machine learning methods connected to the recognition of interruption, malware, and spam.

Keywords : Machine Learning, Deep Learning, Cyber Security, Adversarial Learning

I. INTRODUCTION

With the development of the Internet, cyber-attacks are changing rapidly and the cyber security situation is not optimistic. This survey report describes key literature surveys on machine learning (ML) and deep learning (DL) methods for network analysis of intrusion detection and provides a brief tutorial description of each ML/DL method. Computer systems and web services have become increasingly centralized, and many applications have evolved to serve millions or even billions of users. Entities that become arbiters of information are bigger targets for exploitation, but are also in the perfect position to make use of the data and their user base to achieve better security. Coupled with the advent of powerful data crunching hardware, and the development of more powerful data analysis and machine learning algorithms, there has never been a better time for exploiting the potential of machine learning in security.

Machine learning refers to one aspect of artificial intelligence specifically, to algorithms and processes that “learn” in the sense of being able to generalize

past data and experiences in order to predict future outcomes.

At the most general level, supervised machine learning methods adopt a Bayesian approach to knowledge discovery, using probabilities of previously observed events to infer the probabilities of new events. Unsupervised methods draw abstractions from unlabeled datasets and apply these to new data. Both families of methods can be applied to problems of classification (assigning observations to categories) or regression (predicting numerical properties of an observation). With a supervised method, we would have a set of animals for which we are definitively told their category (e.g. we are given that the dog and elephant are mammals and the alligator and iguana are reptiles). We then try to extract out some features from each of these labelled data points and find similarities in their properties, allowing us to differentiate animals of different classes.

Machine learning algorithms are driven by mathematics and statistics, and the algorithms that discover patterns, correlations, and anomalies in the data vary widely in complexity.

PROBLEM DEFINITION

In this section, we present several issues that must be considered before deciding whether to apply ML algorithms in NOC and SOC. We can anticipate that, at the current state-of-the-art, no algorithm can be considered fully autonomous with no human supervision. We substantiate each issue through experimental results from literature or original experiments performed on large enterprises. We begin by describing the testing environments of our experiments, and the metrics considered for evaluation. The experiments focus on Network Intrusion Detection, and leverage one ML algorithms: K-Means Network Intrusion Detection, we use three labeled real training datasets composed of benign and malicious network flows2 collected in a large organization of nearly 10,000 hosts. The labels are created by flagging as malicious those flows that raised alerts by the enterprise network IDS and reviewed by a domain expert.

II. LITERATURE SURVEY

DeemanYousifMahmood (Classification Trees with Logistic Regression Functions for Network Based Intrusion Detection System)IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 3, Ver. IV (May - June 2017), PP 48-52.

An Intrusion Detection System (IDS) is a system that monitors and analyses network traffics to check for intrusive activities in the network and report events that does not coordinate the security criteria of the system administrator and within the recent years, attacks has been increased rapidly and drastically on networks and web applications which results in a wide interest of researchers for network intrusion detection systems.

IDSs are came into two categories: Signature based and Anomaly based, where the signature based seeks for previously-knowing patterns or samples of attacks, this model can detect and recognize only an attack with a precise matching behavior found versus previously stored patterns or samples knowing by signatures, while anomaly based is build based on origination a profile for normal activity of the system, anomaly based technique promotes itself by comprehension and aggregating information and familiarity about the system and decides the conduct of the security system in view of it.

Battista Biggio, Igino Corona, Blaine Nelson, Benjamin I. P. Rubinstein, DavideMaiorca, Giorgio Fumera, Giorgio Giacinto, and Fabio Roli (Security Evaluation of Support Vector Machines in

Adversarial Environments) Submitted on 30 Jan 2014.

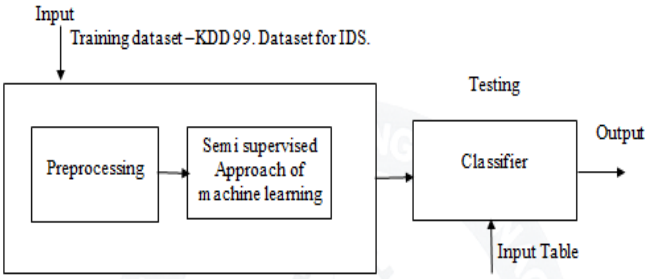
Machine-learning and pattern-recognition techniques are increasingly being adopted in security applications like spam filtering, network intrusion detection, and malware detection due to their ability to generalize, and to potentially detect novel attacks or variants of known ones.

Jiong Zhang and Mohammad Zulkernine (Network Intrusion Detection using Random Forests) 2006 IEEE International Conference on Communications.

With the tremendous growth of network-based services and sensitive information on networks, network security is getting more importance than ever. Although a wide range of security technologies such as information encryption, access control, and intrusion prevention can protect network-based systems, there are still many undetected intrusions. For example, firewalls cannot prevent internal attacks. Thus, Intrusion Detection Systems (IDSs) play a vital role in network security. Network Intrusion Detection Systems (NIDSs) detect attacks by observing various network activities, while Host-based Intrusion Detection Systems (HIDSs) detect intrusions in an individual host.

Yihua Liao, V. RaoVemuri (Using Text Categorization Techniques for Intrusion Detection) 2002 Article. Bibliometrics Data Bibliometrics. · Citation Count: 38 · Downloads (cumulative).

Intrusion detection has played an important role in computer security research. Two general approaches to intrusion detection are currently popular: misuse detection and anomaly detection. In misuse detection, basically a pattern matching method, a user's activities are compared with the known signature patterns of intrusive attacks. Those matched are then labeled as intrusive activities. That is, misuse



detection is essentially a model reference procedure. While misuse detection can be effective in recognizing known intrusion types, it tends to give less.

Trupti A. Kumbhare Prof. Santosh V. Chobe (An Overview of Association Rule Mining Algorithms)Trupti A. Kumbhare et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014, 927-930

Association rule learning searches for relationships among variables. For example a supermarket might gather data about how the customer purchasing the various products. With the help of association rule, the supermarket can identify which products are frequently bought together and this information can be used for marketing purposes. This is sometimes known as market basket analysis. Clustering discovers the groups and structures in the data in some way or another similar way, without using known structures in the data. Classification generalizes known structure to apply to new data. Take an example; an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam" mail.

III. METHODS AND MATERIAL

This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 the Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion

detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

Figure. 3.1. Data Design

KDD

The term Knowledge Discovery in Databases, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

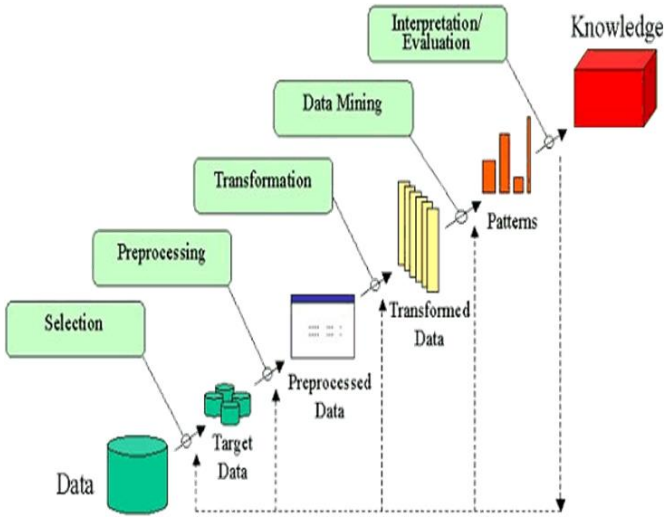


Figure 3.2 : An Outline of the Steps of the KDD **Process**

The following diagram shows the process of knowledge discovery

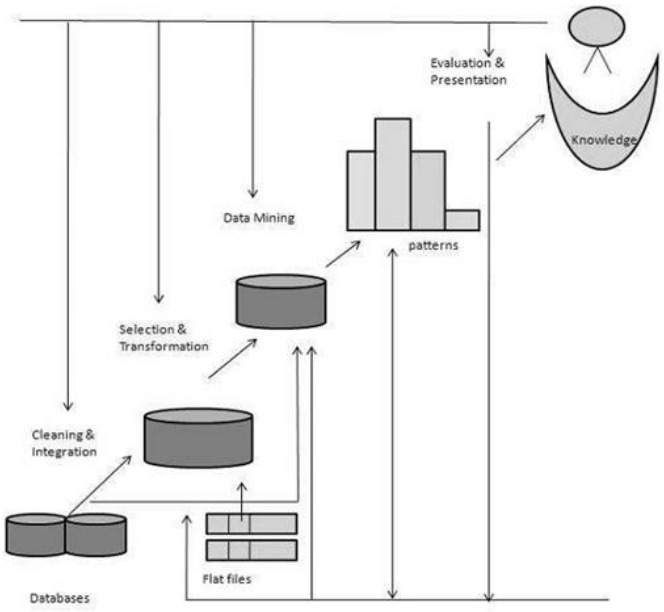


Figure 3.3 : process of knowledge discovery

The KDD CUP 99 Data Set

With the enormous growth of computer networks usage and the huge increase in the number of applications running on top of it, network security is becoming increasingly more important. As it is shown in, all the computer systems suffer from security vulnerabilities which are both technically difficult and economically costly to be solved by the manufacturers. Therefore, the role of Intrusion Detection Systems (IDSs), as special-purpose devices to detect anomalies and attacks in the network, is becoming more important.

K-means clustering algorithm

It is the simplest unsupervised learning algorithm that solves clustering problem's-means algorithm partition n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.

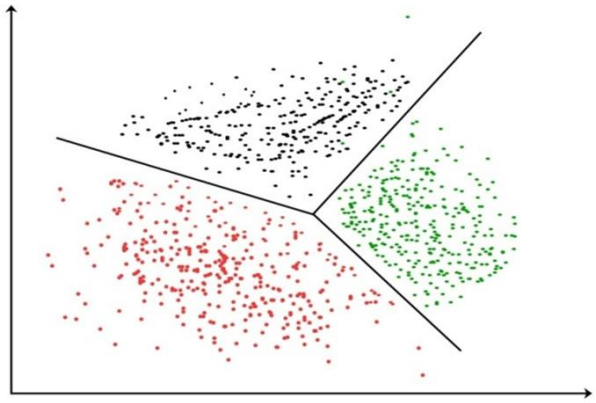


Figure 3.4 : prototype of the cluster

Algorithm:-

1. Specify number of clusters *K*.
2. Initialize centroid by first shuffling the dataset and then randomly selecting *K* data points for the centroid without replacement.
3. Keep iterating until there is no change to the centroid. i.e. assignment of data points to clusters isn't changing.
4. Compute the sum of the squared distance between data points and all centroid.
5. Assign each data point to the closest cluster (centroid).
6. Compute the centroid for the clusters by taking the average of the all data points that belong to each cluster.

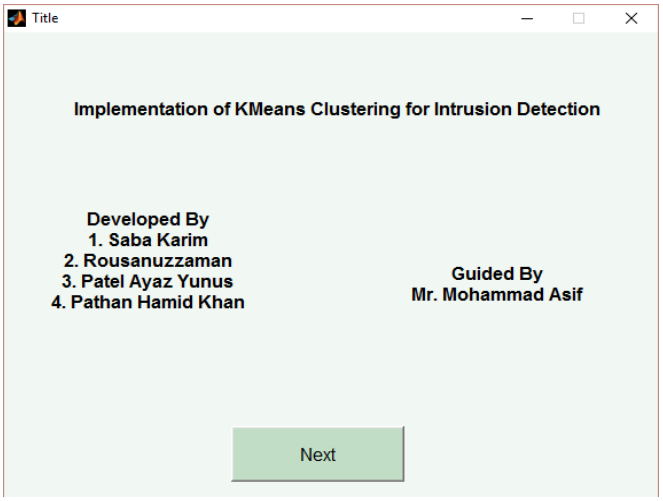


Figure 3.5 : First page after run

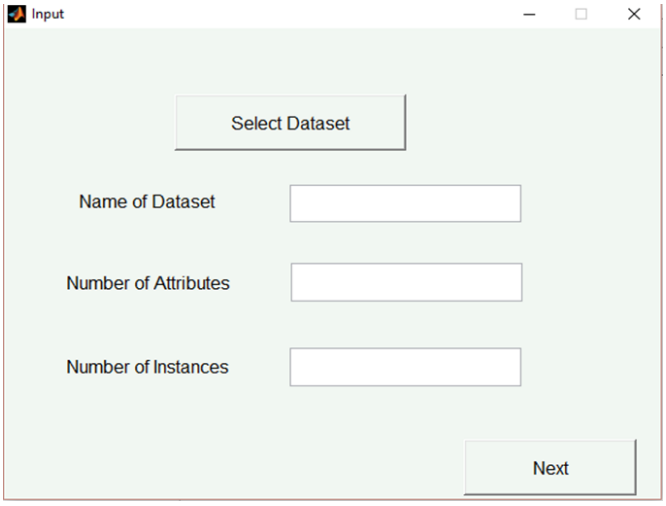


Figure 3.6 : Before selecting the dataset

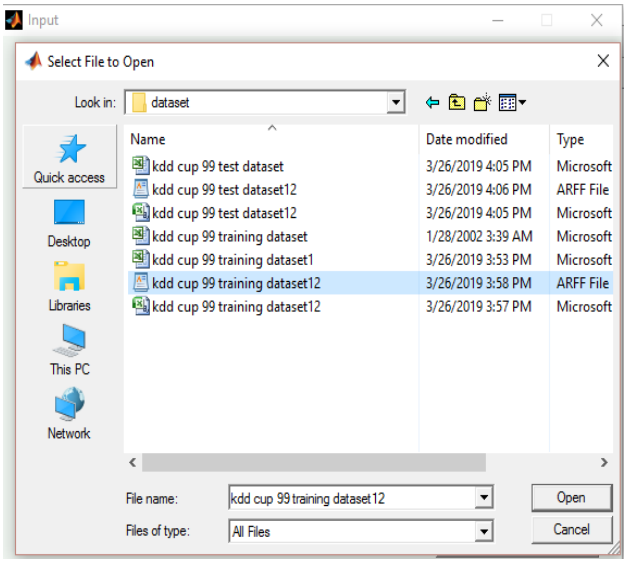


Figure 3.7 : Selecting file for training dataset

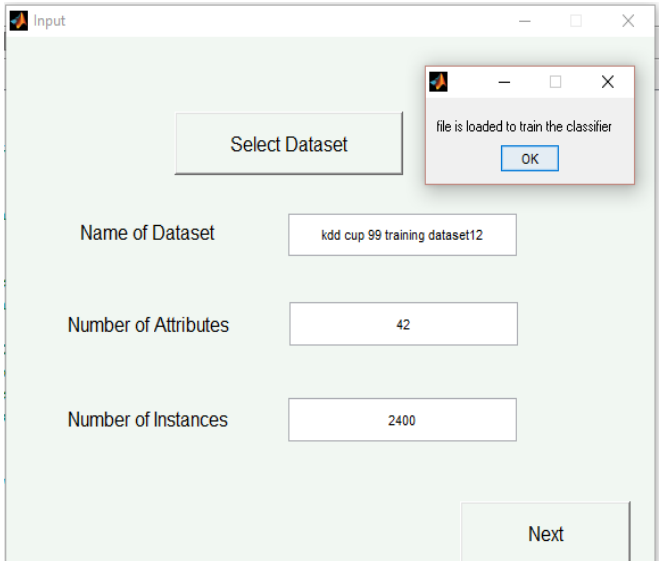


Figure 3.8 : File loaded to train the classifier

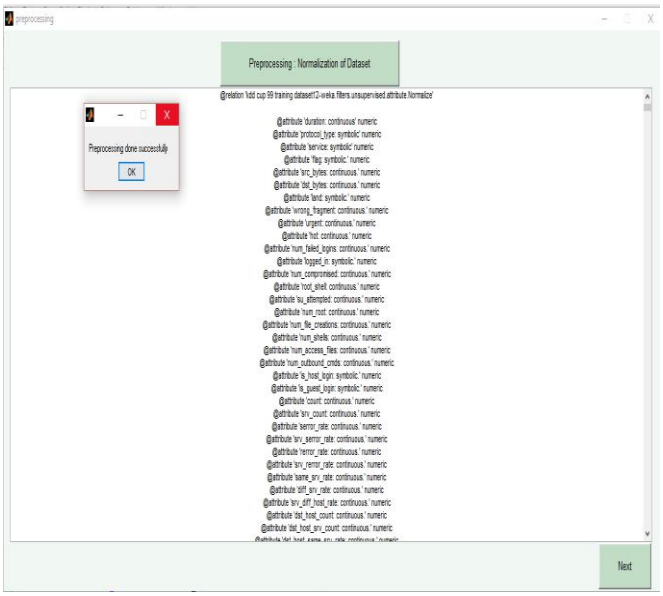


Figure 3.9 : Pre-processing successfully

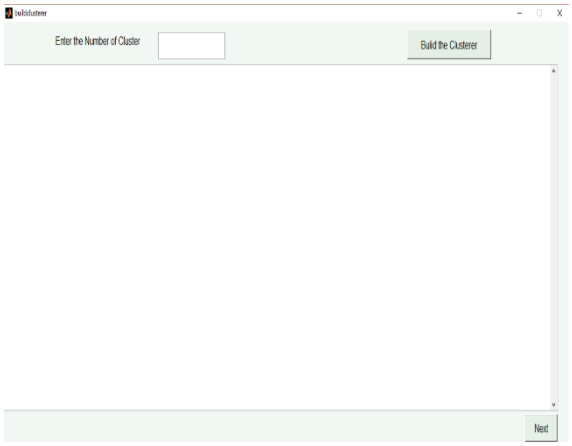


Figure 3.10 : Before entering the number of cluster

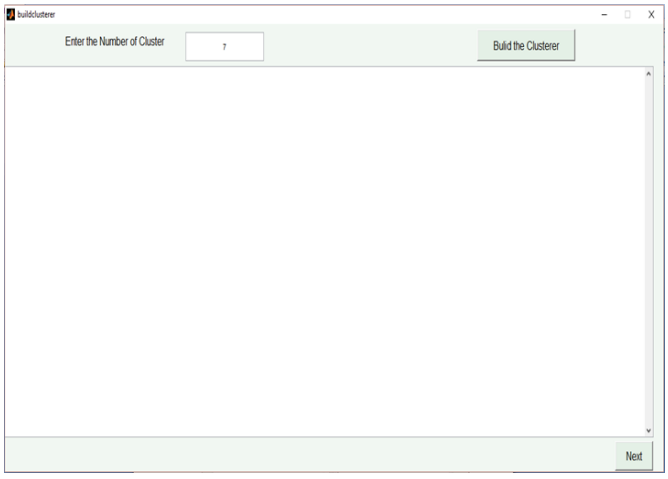


Figure 3.11 : Number of clustered entered 10

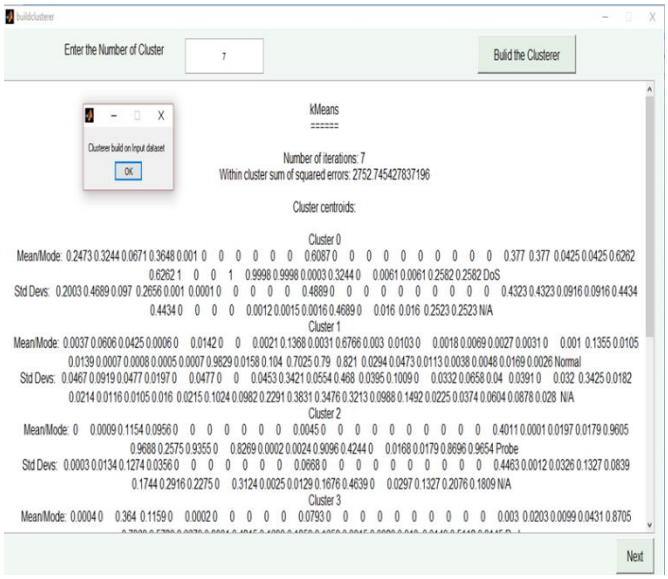


Figure 3.12 : Cluster build on input Cluster

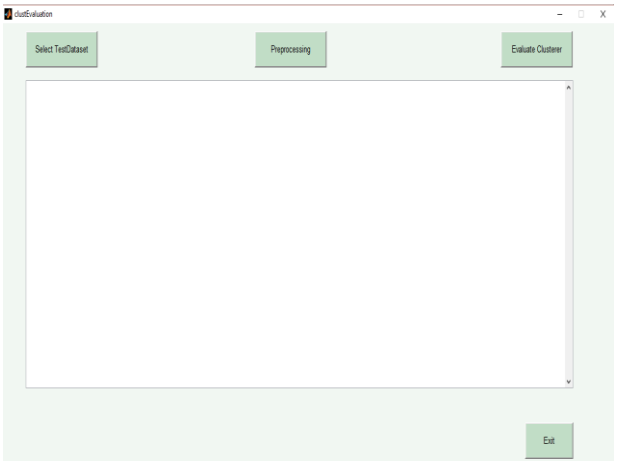


Figure 3.13 : Select Test data

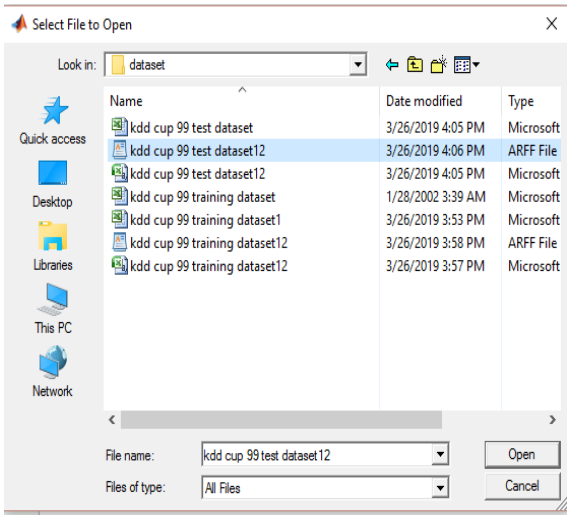


Figure 3.14 : Selecting the Test Dataset

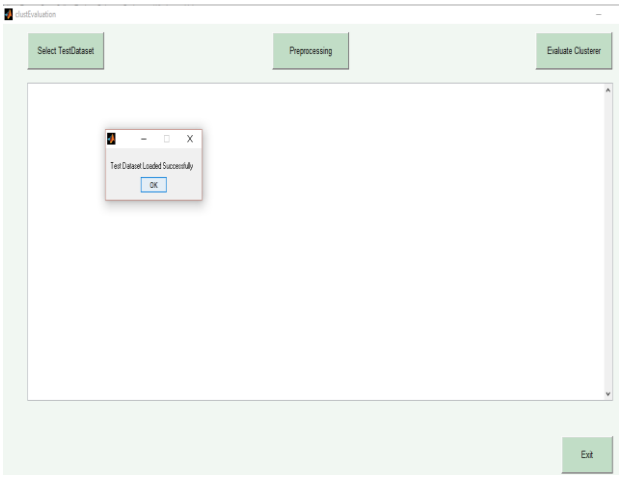


Figure 3.15 : Test Dataset loaded successfully

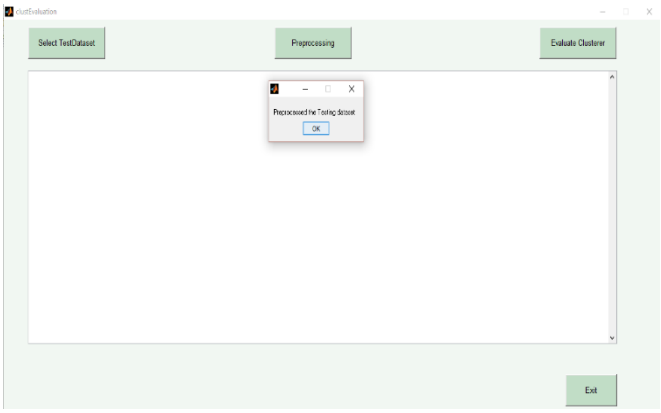


Figure 3.16 : Pre-processed Testing Dataset

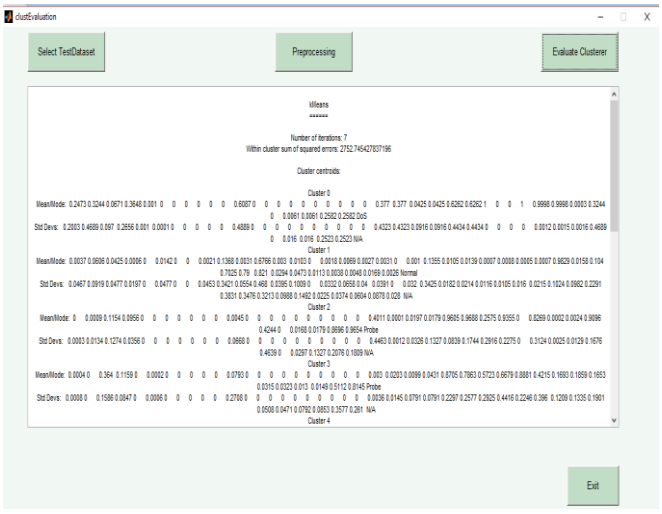


Figure 3.17 : Evaluate Cluster

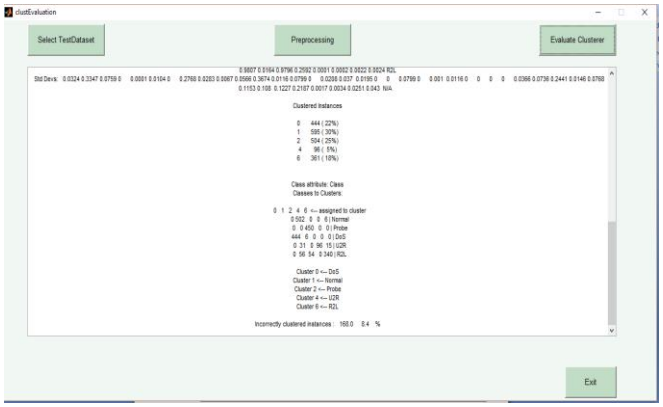


Figure 3.18 : Final output with best efficiency

IV. RESULTS AND DISCUSSION

After selecting the Training dataset we gave a different no of cluster as an input from 1 to 20 and we got the different cluster instances with variant incorrectly instances having different Efficiencies.

But we got the best efficiency with the **7** number of cluster with having:

- ✓ Number of iteration: 7
- ✓ Within cluster sum of squared errors: 2752.745427837196
- ✓ Incorrectly clustered instances :168.0, 8.4 %

Hence, the corrected Cluster instances with efficiency of 91.6 %

We entered the different number of cluster from 1 to 20 ad we got different output which is shown by the following table:

Table 4.1 : Output with different number of cluster

Sr. No.	No. of Cluster	Incorrectly clustered instances	Iteration	In corrected %	Corrected %
1	1	1492	1	74.6	25.4
2	2	1042	13	52.1	47.9
3	3	1042	5	52.1	47.9
4	4	988	6	49.4	50.6
5	5	503	14	25.15	74.25
6	6	501	7	25.05	74.95
7	7	168	7	8.4	91.6
8	8	169	12	8.45	91.55
9	9	384	8	19.2	80.8
10	10	269.0	7	13.45	86.55
11	11	269.0	7	13.45	86.55
12	12	415.0	17	20.75	79.25
13	13	433.0	15	21.65	78.35
14	14	434.0	15	21.7	78.3
15	15	457.0	12	22.85	77.15
16	16	477.0	12	23.85	76.15
17	17	494.0	11	24.7	75.3
18	18	494.0	11	24.7	75.3
19	19	574.0	11	28.7	71.3
20	20	586	11	29.3	70.7

V. CONCLUSION AND FUTURE SCOP

The autonomous capabilities of ML algorithms must not be overestimated, because the absence of human supervision can further facilitate skilled attacker to infiltrate, steal data, and even sabotage enterprise. The machine learning is a very vast field of computer science in modern technology, through the availability of internet. The major issues in this evolutionary world are to communicate data passes data deal with security.

Machine and deep learning approaches are increasingly employed for multiple applications and are being adopted for cyber security, hence it is important to evaluate when and which category of algorithms can achieve adequate results. We Analyses these techniques for three relevant cyber security problems: intrusion detection malware analysis and spam detection.

Our results provide evidence that present machine learning techniques are still affected by several shortcomings that reduce their effectiveness for cyber security.

As a possible future development to the implementation of the proposed system, one can include more attack scenarios in the dataset and used latest dataset.

Our future work involves development of intrusion protection system to achieve low false positive rate and more accuracy-using anomaly based detection approach.

VI. REFERENCES

[1]. S. Aftergood, ``Cybersecurity: The cold war online,'' Nature, vol. 547,no. 7661, pp. 30_31, Jul. 2017.

[2]. M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” Science, 2015.

[3]. A. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” IEEE Communications Surveys & Tutorials, 2015.

[4]. E. Blanzieri and A. Bryl, “A survey of learning-based techniques of email spam filtering,” Artificial Intelligence Review, 2008.

[5]. J. Gardiner and S. Nagaraja, “On the Security of Machine Learning in Malware C8C Detection,” ACM Computing Surveys, 2016.

[6]. DeemanYousifMahmood (Classification Trees with Logistic Regression Functions for Network Based Intrusion Detection System)IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 3, Ver. IV (May - June 2017), PP 48-52.

[7]. Battista Biggio, Igino Corona, Blaine Nelson, Benjamin I. P. Rubinstein, DavideMaiorca, Giorgio Fumera, Giorgio Giacinto, and Fabio Roli (Security Evaluation of Support Vector Machines in Adversarial Environments) Submitted on 30 Jan 2014.

[8]. Jiong Zhang and Mohammad Zulkernine (Network Intrusion Detection using Random Forests) 2006 IEEE International Conference on Communications.

[9]. Yihua Liao, V. RaoVemuri (Using Text Categorization Techniques for Intrusion Detection) 2002 Article. Bibliometrics Data Bibliometrics. · Citation Count: 38 · Downloads (cumulative).

[10]. Trupti A. Kumbhare Prof. Santosh V. Chobe (An Overview of Association Rule Mining Algorithms)Trupti A. Kumbhare et al, /

(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014, 927-930

[11]. On the Effectiveness of Machine and Deep Learning for Cyber Security 2018 10th International Conference on Cyber Conflict CyCon X: Maximising Effects T. Minárik, R. Jakschis, L. Lindström (Eds.) 2018 © NATO CCD COE Publications, Tallinn.

[12]. Graphical User Interface for Intrusion Detection in Telecommunications Networks 28 March 2011.

[13]. Prakash Ranganathan, Juan Li, Kendall Nygard, "A Multiagent System using Associate Rule Mining (ARM), a collaborative filtering approach", IEEE 2010, pp- v7 574- 578.

[14]. S.Devaraju, S.Ramakrishnan:,"Analysis of Intrusion Detection System Using Various Neural Network classifiers, IEEE 2011.

[15]. Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, K. Dai, "Anefficient intrusion detection system based on support vector machines and gradually feature removal method", Expert Systems with Applications,2011,39: p. 424 430.

[16]. A.M. Chandrasekhar, "Intrusion Detection Technique By Using K-Means, Fuzzy Neural And Svm Classifier ",2013 International Conference on Computer Communication and Informatics (ICCCI - 2013), Jan 04-06, 2013 Coimbatore, India.

[17]. Hesham Altwaijry, "Bayesian Based Intrusion Detection System ", Journal of King Saud University–Computer and Information Sciences (2012) 24,1–6.

[18]. V. B. Kosamkar and S. S. Chaudhari "Data Mining Algorithms for Intrusion Detection System: An Overview". IJCA Proceedings on International Conference on Recent Trends in Information Technology and Computer Science 2012 ICRTITCS(3):9-15, February2013.

[19]. S. Duque and M. N. B. Omar "Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)". Proceedings of Science direct: Procedia Computer Science 61, pp. (46-51), 2015.

[20]. D. Ndumiyana, R. Gotora and H. Chikwiriro "Data Mining Techniques in Intrusion Detection: Tightening Network Security". International Journal of Engineering Research and Technology, Vol. 2 Issue 5, IJERT, May 2013.

[21]. M. K. Asif, T.A. Khan, T.A. Taj,U.Naeem and S. Yakoob "Network Intrusion Detection and its Strategic Importance". Business Engineering and Industrial Applications Colloquium (BEIAC), IEEE, 2013.

[22]. A. Bohara, U. Thakore and W. H. Sanders "Intrusion Detection in Enterprise Systems by Combining and Clustering Diverse Monitor Data". Department of Computer Science University of Illinois at Urbana-Champaign. Proceeding HotSos '16 Proceedings of the Symposium and Bootcamp on the Science of Security, Pages 7-16, Pittsburgh, Pennsylvania— April 19 - 21, 2016.

[23]. M. Mandanna, L. Kiran and R. P. Madhavi "Implementation of Intrusion Detection Using Genetic K-Means Algorithm in Wireless Sensor Networks".Dept. of CSE BMSCE Bangalore, India, International Journal of Advance Research in Computer Science and Management Studies, Volume 4, Issue 3, March 2016.

[24]. J. Jabez and B. Muthukumar "Intrusion Detection System (IDS): Anomaly Detection using Outlier Detection Approach". Sathyabama Unicersity, Sholinganallur, Chennai, International Conference on IntelligentComputing, Communication and Convergence (ICCC-2015), Procedia Computer Science 48:338-346, December 2015.

[25]. A. P. Beldar andV. S.Wadne "The Detail Survey of Anomaly/Outlier Detection Methods in Data Mining". International Journal of

Multidisciplinary and Current Research, Vol.3, May/June 2015 issue.

[26]. P. S. Rath, M. Hohanty, S. Acharya and M. Aich, "Optimization of IDS Algorithms Using Data Mining Technique", Proceeding of 53rd IRF International Conference, Pune, India,ISBN 978-93-86083-01-2, 2016.

[27]. Md.E. Haque and T.M. Alkharobi, "Adaptive Hybrid Model for Network Intrusion Detection and Comparison among Machine Learning Algorithms", International Journal.

[28]. M. Dhakar and A. Tiwari, "A Novel Data Mining based Hybrid Intrusion Detection Framework", Journal of Information and Computing Science, 2014, Vol-9 No-1 pp. 037-048, ISSN 1746-7659, England, UK..

[29]. TR. Patel, A. Thakkar and A. Ganatra, "A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems", International Journal of Soft Computing and Engineering (IUSCE), March-2012, Vol-2, Issue-1, ISSN: 2231-2307.

[30]. Somani Manish and Roshni Dubey, "Hybrid Intrusion Detection Model Based on Clustering and Association", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.3, Issue 3, ISSN (Print): 2320-3765, ISSN(Online):2278-8875 March 2014.

[31]. R. Venkatesan, R. Ganesan and A.A.L. Selvakumar, "A Comprehensive Study in Data Mining Frameworks for Intrusion Detection", International Journal of Advanced Computer Research, December-2012, Volume-2 Number-4 Issue-7, ISSN (print): 2249-7277 ISSN (online): 2277-7970.

[32]. Heba F. Eid et al., "Principal Components Analysis and Support Vector Machine based Intrusion Detection System", 10 th International Conference on Intelligent Systems Design and Applications, (IEEE, 2010).

Cite this article as :

Saba Karim, Rousanuzzaman, Patel Ayaz Yunus, Patha Hamid Khan, Mohammad Asif, "Implementation of K-Means Clustering for Intrusion Detection", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 1232-1241, March-April 2019. Available at doi :
https://doi.org/10.32628/CSEIT1952332
Journal URL : http://ijsrcseit.com/CSEIT1952332