International Journal of Engineering & Technology, 7 (4.19) (2018) 600-604

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Anomaly Detection System for Internet Traffic based on TF-IDF and BFR Clustering Algorithms

Suad A. Alasadi*[1], Wesam S. Bhaya[2]

College of Information Technology, University of Babylon, Babil, Iraq.
Corresponding Author E-mail: it.suad.abdulelah@uobabylon.e

Abstract

An anomaly can be defined as any deviation from the normal and something which is outside the usual range of variations, it consumes network resources, and lead to security issues such as Confidentiality, Integrity, and Availability (CIA). An Intrusion Detection Systems (IDS) are designed and implemented by many researchers to analyze, detect, and prevent the anomaliestraffics. Although, there are various techniques for IDS to detect anomalies like statistical, machine learning techniques. Data mining can be efficiently employed for anomaly detection. Since, it works to extract features from network traffic; it can be used to distinguish between common legitimate and attack traffics. Data mining can be efficiently identifying the important data for user and predicts the results that can be utilized to detect various types of attacks.

In this paper, an anomaly detection approach usingTerm Frequency Inverse Document Frequency(TF_IDF) and Bradley, Fayyad, and Reina(BFR) clustering algorithm is presented to detect and prevent malicious traffic efficiently and with low time complexity.Multiple types of attacks are detected in the proposed solution like (Flooding, Denial of Service (DoS), Backdoors, and Worms)attacks effectively using two modern datasets are which are"NUST2009, UNSW-NB2015".

The experiments result shows that the BFR clustering algorithm perform better than the K-meanalgorithm in term of accuracy and detection rate. The overall accuracy for NUST2009 dataset is 99.2%, the detection rate is 100%, and false alarm rate is 0%. While the overall accuracy in UNSW-NB2015 dataset is 98.76, the detection rate is 79.28%, and false alarm rate is 0%.

Keywords: *Anomaly Detection,IDS, Network Attacks, Clustering Data Mining, TF_IDF, BFR.*

1. Introduction

The increasing numbers of malicious threats on the Internet and computers networks are commonly shown in Internet traffics, they may be Denial of Service attacks(DoS),worm, flooding attack. These attacks consume network resources, and led to prevent the legitimate users from accessing to network resources. Thus, detecting such anomalies accurately has become an important problem for the network community to solve[1].

IDS's have two main categories: Signature-based detection works by monitoring network packets and comparing it with a database of signatures or patterns from known malicious attacks. Usually signatures is a combination of packet header and packet content to determine the anomalous traffic flows. Anomaly based detection in the other hand, works by monitoring packets on the network and comparing it with a baseline profile of normal packets. This profile recognize what is normal for the network such as: the normal bandwidth usage, the common protocols used and detect the anomaly as any packet which is different from the normal profile [2].

A very common example of network attacks is flooding attack. Flooding attacks have become a serious attack, because they do not only lead to a loss of confidence and privacy, but also to illegal actions taken against an organization. Flooding attacks work by sending useless packets from different sources to the target in a short period of time, which led to consume the resources, making them unavailable for normal operations. There are three well-known flooding attackswhich are User Datagram Protocol(UDP), Internet Message Control Protocol(ICMP), and Transmission Control Protocol/ synchronize(TCP SYN) attacks [3].Figure (1) illustrated the TCP SYN flooding attack.

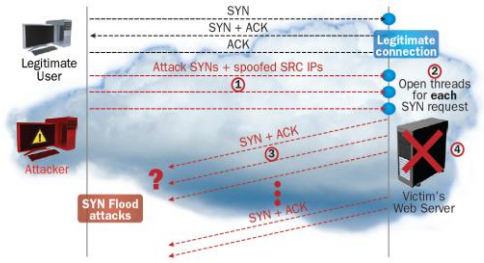

Fig. 1: TCP SYN Flooding Attack.

As shown in this figure, the TCP connection includes three-way handshake, an attacker will send enormous number of SYN packets without send any ACK which make the server wait for missing ACK's. If the server has only incomplete buffer line for new connection, this attack will make the server incapable to manner other received connection due to the overloaded of the queue [4].

Data Mining (DM) is the process of extraction useful patterns and models from a huge dataset. DM can be used to analysis network traffic Data mining can be used for anomaly detection. Since, it works to extract features from network traffic; it can be used to

distinguish between common legitimate and attack traffics. Data mining includes several techniques which may be supervised or unsupervised. The supervised techniques depend on label to train the algorithm, while the unsupervised techniques group the similar objects based on particular measures and don't have label class to earn the algorithm. Figure 2 shows data mining techniques [5].

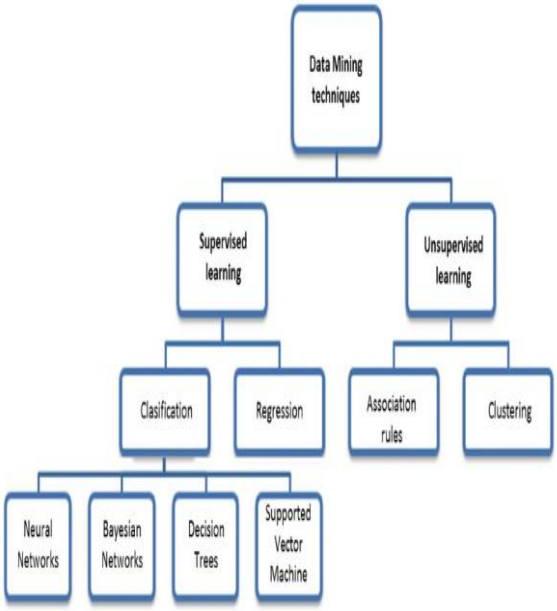

Fig. 2: Data mining Techniques

In this paper, the unsupervised clustering method (BFR) has been implemented toanalyze and detect multiple types of attacks.The other parts of this paper are organized as section2 shows the related works, section3: presents network flow datasets, Section4: explains anomaly detection system, section5 shows system evaluation. Finally, section 6 explains systemresults.

2. Related Works

Xin Du, Yingjie et al. used K-means clustering data mining technique for intrusion detection. They collected data set from MIB network different time interval. The authors applied information entropy to select most significant attributes from the entire set of data set. They concluded that the approach has given good performance to select better featureattributes and high accuracy detection rate[6].

FarhadSoleimanianet al. introduced K-means and Fuzzy K-means data mining clustering approaches for intrusion detection system. They collected data set from KDD cup which has 41 attributes. They use k-means and fuzzy k-means approaches to identify the type of DoS attacks. They concluded that the fuzzy k-means method achieved slightly better than k-meansmethod in detecting denial of service (DoS) attack[7].

Miller, W.Deritrik, W.Hu et al. proposed Den Stream and frequency histogram approaches for detecting anomalous packets based on data stream mining. Authors collected network data from different database DARPA and MCPAD data set. The Den Stream approach treat individual packets as points and are flagged as normal or malicious based on whether these points are normal and outliers. They utilized a histogram approach to buildthe histogram model for new packet payload. They used Pearson correlation for computing between twohistograms. From the result, they concluded that histogram-based detection algorithm achieved little better performance but required more numbers of features than the clustering-based algorithm[8].

Ghanshyam Prasad Dubey et al. proposed RST and incremental SVM approaches to detect intrusion. According to the authors, the incremental SVM approach increased performance for intrusion detection. They noted the RST and increment SVM approaches are effective to decrease the space density of data[9].

Shailendra Kumar et al. introduced rough set theory and support vector machine for dimensionality reduction in intrusion detection. They experimented with KDD cup data set. They applied rough set theory to select the most significant attributes from KDD cup data set. A comparative analysis between SVM with original 41 data set and reduction data set is presented. They concluded that theproposed algorithm is very reliable for intrusion detection[10].

Heba F. Eid et al.analyzed intrusion detection system using Support Vector Machines with Principal Component Analysis approaches. They tested their proposed model on NSL-KDD data set. The PCA approach used to reduce the number of features in order to decrease the complexity of the system. Their results showed that the proposed system is capable to speed up the process of intrusion detection and to reduce the memory space and CPU time cost[11].

Vivek K. Kshirsagar et al. proposed decision tree data mining techniques for investigating and evaluating intrusion detection. The authors collected data set from DARPA which has a different type of attacks. The SVM gave better result than decision tree on DoS class. They observed that the decision tree is more efficient for detecting the intrusion[12].

S. Mehibs and S. Hashim,proposed a network intrusion detection system Based on fuzzyc_mean algorithm in cloud computing environment [13]. In addition, they propose a network intrusion detection system in a cloud environment based on backpropagation neural network [14].

Bhaya, W. and Ebadymanaa, M.combinedunsupervised data mining techniques as intrusion detection system. The entropy concept in term of windowing the incoming packets is applied with data mining technique using Clustering Using Representative (CURE) as cluster analysis to detect the DDoS attack in network flow [15].

3. Network Flow Datasets

Network traffic analysis has become more and more vital and important in present day for monitoring intrusion attack [16]. One of the main research challenges for intrusion detection is the unavailability of network datasets which can mirror modern network traffic scenarios. In the proposed system, two modern benchmark datasets are used:

3.1 NUST2009 Dataset

Each record of the dataset represents a packet in a 10 features format. Dataset packets are collected from campus network. It consists of 1,000,000 packets from normal and flooding attack types [17].

3.2 UNSWNB15 Dataset

This effort is made by the cyber security research group at the Australian Centre for Cyber Security (ACCS) and other researchers. It is used to evaluate NIDSs. This dataset has 2, 540,044 records stored in CSV files. Furthermore, a part from this dataset was separated into a train and test sets. The training set includes175,341 records, while the testing set had 82,332 records from all complicated attacks and normal records. It contains ten different classes, one for normal and nine types for security events and malware [18].

4. Proposed Anomaly Detection System

In this paper, a proposed system called "Anomaly Detection Based on BFR clustering" is described to detect anomalies packets. The proposed system includes three main phases: (1) the preprocessing phase which improves the quality of the network traffic; (2) the anomaly detection phase which used to detect the

attacks; (3) the evaluation phase which includes the performance measures to evaluate the results.

4.1. Preprocessing Stage

Multiple types of attacks and normal traffic are collected from the two datasets (NUST2009, UNSW-NB1) as described in section3. Two million (2,000,000) network packets are selected from both data sets for attack and normal traffic. The proposal system works on TCP/IP header information of the TCP/IP packets. The most important features (attributes) are described in Table (1).

Table 1: NUST2009 data set features

Name	Description
TimeStamp	received packet time
PacketSize	actual packet length in byte
SourceIP	the network layer source IP address
Destination IP	the network layer destination IP address
Source Port	the transport layer source port
Destination Port	the transport layer destination port
TCP Flags	four important TCP flags
Transport Protocol	the transport layer protocol integer code
Packet Direction	direction of the packet
Packet Type	normal or attack

4.2. Data Transformation Using TF-IDF Weight Factor

It includes transforming the data to forms suitable for the mining process. Term frequency-inverse document frequency (TF-IDF) method is used in this pre-processing step. It is used to convert data from nominal data to numeric data. It is a data mining method which enables to determine a weight for each packetin the dataset [19]. The proposed system is the first that employed TF.IDF for intrusion detection system. This method helped in analyzing large network traffic, early detection of an attack and improving both accuracy and processing speed for the system.If f_{ij} the occurrence number of word i in document j. Then, the term frequency TFijare defined as:

$$TFij = \frac{fij}{maxkfkj} \quad ..(1)$$

f_{ij}: the term frequency of term i in document j is divided by the maximum number of the frequency of any term in the same document.

If termi looks in n_i of the N total documents. Then

$$IDFi = \log2(N / n_i) \quad \ldots\ldots \quad . \quad (2)$$

Then $TF_{ij} \times IDF_i$ will be the weight for word i in document j. The terms with the highest TF.IDF score is always the terms thatrepresent the topic of the document.

Algorithm (1) showsTF-IDFdata transformation algorithm.

Algorithm (1): Pre-processing of Network Traffic using TF.IDF

Input: NT (Network Traffic) consist of (f) features, Document length (doclen).
Output: ***TF.IDF_List***(List of **TF.IDF** values)
1. Begin
2. Define class Packet of Network Traffic (NT) with features (Psize, SIP, DIP,….. number of selected features)
3. Read network data (packets) from database (DS) with f-feature
4. Using Sliding Window to get N Documents (Doc)with (doclen) size
5. Set doclen← 50 **(for NUST, UNSWNB15)**
6. Set j ← 0
7. Repeat {
8. If (DWj. Size () <doclen)
9. Docj.add (Psize, SIP, DIP, Sport, Dport … number of features)
10. else
11. {
14. For each feature in Docj do {
15. w**ij ← Calculate** frequency for each item in **distinct feature** (Grouping psize, source IP, dest. IP ...etc.) and save it in dictionary (**K**, **V**), where **k**: the term I, **V**: count of termi in docj
16. **C**alculate **TFfor** all items in Doci**u**sing**equation (1)** andsave it in dictionary (k,tf**ij**)
17. max ← **Calculate** maximumfrequency **of** an item**i** in Doc**j**
18. Tf**ij** ← w**ij** / max
19. ***TF_List.add (***Tfij) *// TF_List is a list of f-dimensional features*
20. }
21. j ← j + doclen;
22. }
23. } Until (End of Size DS)
24. Repeat{
25. For each feature in Docj do {
26. Define class NumOfAppearsPerDocument to count the occurrence of an item in the total documents(ni)
27. Calculate the **IDF** for all items in Docj using **equation (2)**
28. Dfij = log2(N / ni)
29. Calculate **TF.IDF** for all items in Doci
30. TFDFij= Tfij * Dfij
31. ***TF.IDF_List.add*** (TFDFij)
32. }
33. j ← j + doclen
34. } Until (End of Size DS)
35. End Algorithm

4.3 Bradley, Fayyad, and Reina (BFR) Clustering Algorithm

This algorithm is named BFR to its authors in the1998[65]. It is a point-assignment clustering algorithm. It is an extension for K-mean clustering algorithm to cluster large scale datasets in a high-dimension space. It has a strong hypothesis about the shape of clusters: they must be normally spread about a centroid. The number of clusters is known (k). Firstly, it selects the initial k points in the same way of k-mean selection. Then, the points of the dataset are read in chunks.

This algorithm clusters the data points in three different sizes sets. These sets are:

1. Discard Set: it represents the summaries of each cluster. The summaries of clusters are not "discarded"; they are in fact very important. However, the points which are represented by the summary are discarded.
2. Compressed Set: These are summaries of set of points close one another like to the cluster summaries, but not close to any cluster. It is called miniclusters.
3. Retained Set: it set of points that cannot assign to any cluster nor to any point in a compressed set.

In the proposed system, the dataset is divided into training and testing data, training data are used to train the model. These data are inputs to the BFR algorithm to produce pattern information. The pattern information's are the summarizes for each cluster which is discussed above. Testing data used pattern information to test the system performance.Algorithm (2) illustrates overall steps of clusteringalgorithm.

Algorithm (2): BFR Clustering Algorithm

Input:Train_data (**DS**)//Each packets*P* consists of multi f features
Output: **CC** (list of Clusters centroids), clusters summarization (**SUMi** , **SUMSQi**) .
1. Begin
2. Choose **N** clusters **// N=4 in datase1, N=5 in dataset2**
3. for i= 1 to N{
4. CC_i← Choose centroids of N clusters from **tranning_list** randomly
5. Assign each training object p_j to closest centroids (cC_i)
6.foreach p in **train_data**
7. min-Euclidean-distance ← $\sqrt{\sum_{j=1}^{f} (pj - CCj)\,2}$
8.}
9. Update centroids (c_i) after assign each point to the closest clusters
10. for i=1 to N do
11.foreach point p_j in C_i do

12. C_i_mean = $\frac{\sum_{i=0}^{C} pj}{m}$ // m is the total number points in the cluster
13. }
14. **Compute SUM, SUMSQ** for each cluster
15. **Foreach cluster Ci do{**
16. **For (a=0; a<C_i.Count){**
17. **SUM_i +** ← p_a
18. **$SUMSQ_i$** + ← $(p_a)^2$
19. **}**
20. **}**
21. Compute variance v, standard deviation sd for each cluster
22. **Foreach cluster Ci do**
23. **For (a=0; a<C_i.Count) {**
24. $v_i \leftarrow \frac{SUMSQi}{Ci.Count}$
25. $sd_i \leftarrow \sqrt{variance}$
26. **}**
27. **}**
28. **End.**

5. Evaluation Method

The performance measures for intrusion detection can be calculated by a confusion matrix [21]. Confusion Matrix: is used to summarize the predictive performance of a classifier on the test data. It is common in a two-class format but can be generated for any number of classes [22]. A confusion matrix for two classes is shown in Table 2:

Table 2: Confusion Matrix

Actual	Predicted	
	Normal	Attack
Normal	TN	FP
Attack	FN	TP

True Positive (TP): Number of the instances that are correctly classified as an attack.
True Negative (TN): Number of the instances that are correctly classified as normal.
False Positive (FP): Number of normal instances that are were classified as an attack.
False Negative (FN): Number of attack instances that are were classified as normal.
To evaluate the results, we have used standard metrics such as detection rate DR, false alarm FA, and accuracy.

$$\text{Detection rate} = \frac{TP}{TP + FN} \quad \ldots (3)$$

$$\text{False alarm} = \frac{FP}{FP + TN} \quad \ldots (4)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \ldots (5)$$

A successful anomaly detection algorithm should achieve high DR, high accuracy and low FA.

6. Experimental Results

Clustering results of NUST2009, UNSW-NB15 Datasetsare shown in tables (3)(4) using a confusion matrix. Tables (5) (6) show the results in term of accuracy, detectionrate and false alarm rate based on BFR method. The results are calculated for testing data based on formulas in Equations (3,4,5) respectively.

Table 3: Confusion Matrix for Testing Phase for NUST2009 Dataset

Actual	Predicted	
	Normal	Attack
Normal	TN(1768)	FP(44)
Attack	FN(0)	TP(2688)

Table 4: Confusion Matrix for Testing Phase for UNSW-NB15 Dataset

Actual	Predicted	
	Normal	Attack
Normal	TN(524)	FP(11)
Attack	FN(0)	TP(352)

Table 5: BFR Performance Evaluation using NUST2009

BFR values	Performance Measure				Predicted clusters
	Accuracy %	Detection Rate	(FAR) %	F. Measure %	
4	99.02%	100%	0	95.56%	Normal+ (TCP+UDP+ICMP) flood cluster

Table 6: BFR Performance Evaluation using UNSW-NB15

BFR values	Performance Measure				Predicted clusters
	Accuracy %	Detection Rate	False Alarm Rate (FAR) %	F. Measure %	
5	98.76%	100%	0	79.28%	Normal+ (Fuzzers, DoS, Exploits, Generic)

Figures (3) (4) shows the spherical shapes for discard points clusters of NUST2009 dataset.

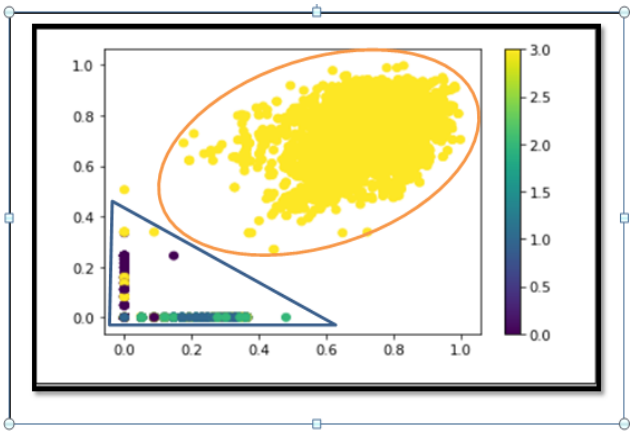

Fig. 3: Two Spherical shape clusters for NUST Dataset Discard Points

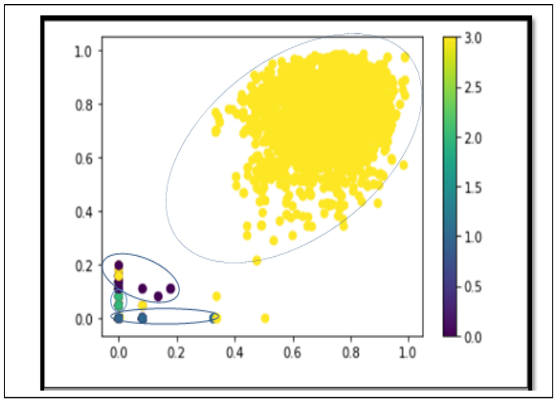

Fig. 4: Four Spherical shape clusters for NUST

Dataset Discard Points

As shown in figure (3), the discard points are grouped into two clusters normal and attack points, and in figure (4) shows that attack points are more specific grouped into multiple types.
Table (7) presents the comparison of the proposed system with other works which used UNSW-NB15 dataset, in term of the accuracy, false alarm, detection rate, and F. Measure

Table 7: Comparison between a Proposed Detection Model and Other Models

No.	Method	Accuracy	DR	FAR	F.Measure	Dataset
1.	Statistical model(BMM) 2017	93.4%	92.7 %	5.9%		UNSW-NB15
2.	Statistical model GAA-ADS 2017	91.8%	91.0 %	5.8%	———	UNSW-NB15
3.	TF.IDF_BFR Proposed System	98.76%	100%	0.02 %	79.28%	UNSW-NB15

Figure (5) shows results comparison of the proposed system with other works based on accuracy, DR, FAR for UNSW-NB15.

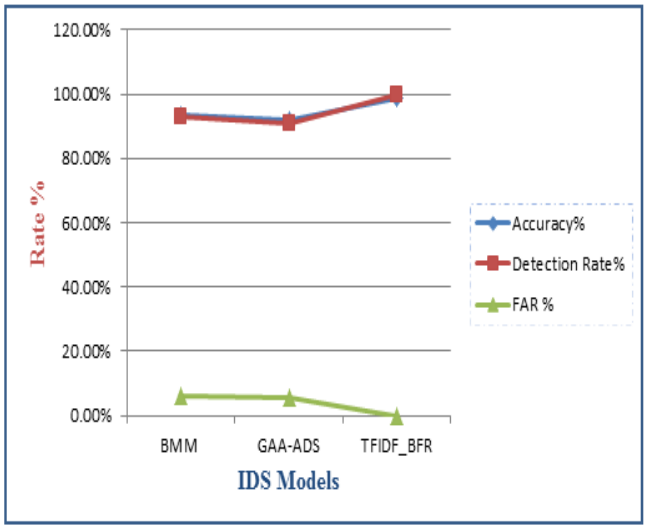

Fig. 5: Results Comparison of Proposed System with Other Works

7. Conclusion

This paper proposed ways to detect and identify harmful packets in a network that consumes network resources and prevents actual users from accessing the network and violating network security.The proposed system includes several methods of data mining to detect several types of attacks which violate the network. The proposed work which employs TF_IDF and BFR clustering algorithm are not previously used in scientific research to detect network attacks.The proposaldetects multiple types of attacks with low time complexity, and we achieved excellent results on very large network data efficiently.The overall accuracy for NUST2009 dataset is 99.51%, the detection rate is 100%, false alarm rate is 0.01%, and F. measure is 96.32%. While the overall accuracy in UNSW-NB15 dataset is 98.76, the detection rate is 100%, false alarm rate is 0.02%, and F. measure is 79.28%. This done with spherical shape's clusters and in an efficient time (in seconds).

References

[1] Marnerides, A. Schaeffer-Filho, and A. Mauthe, “Traffic anomaly diagnosis in Internet backbone networks: A survey,” Elsevier, vol. 73, pp. 224–243, 2014.

[2] S. Kumar, “Survey of Current Network Intrusion Detection Techniques,” Citeseer, pp. 1–18, 2007.

[3] C. Douligeris, A. Mitrokotsa, “DDoS Attacks and Defense Mechanisms: Classification and State-of-the-Art”, Computer Networks, Vol. 44, No. 5, pp. 643-666, 2004.

[4] R. kumar, M. Nene," A Survey on Latest DoS Attacks: Classification and Defense Mechanisms", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 8, October 2013.

[5] A. Rajaraman and J. D. Ullman, “Mining of Massive Datasets,” Lect. Notes Stanford CS345A Web Mining, vol. 67, p. 328, 2011.

[6] Xin Du, Yingjie Yang, Xiaowen Kang, “Research of Applying Information Entropy and Clustering Techniques Network Traffic Analysis”, IEEE, 978-0-7695-3508-1, 2008.

[7] Farhad S. Gharehchopogh, NedaJabbari, and Zeinab G. Azar. “Evaluation of fuzzy k-means and k-means clustering algorithms in intrusion detection systems”. International Journal of Scientific and Technology Research, 1(11) 66–71, December 2012.

[8] Z. Miller, W. Deitrick, and W. Hu, “Anomalous Network Packet Detection Using Data Stream Mining,” J. Inf. Secur., vol. 2, no. 4, pp. 158–168, 2011.

[9] Ghanshyam P. Dubey, Neetesh Gupta, and Rakesh K. Bhujade. “A novel approach to intrusion detection system using rough set theory and incremental svm”. International Journal of Soft Computing and Engineering (IJSCE), (1):663–667, 2011.

[10] R.-C. Chen, K.-F. Cheng, Y.-H. Chen, and C.-F. Hsieh, “Using Rough Set and Support Vector Machine for Network Intrusion Detection System,” in 2009 First Asian Conference on Intelligent Information and Database Systems, 2009, pp. 465–470.

[11] Eid H. F., Darwish A., Ella Hassanien, and Abraham A. “Principle components analysis and support vector machine based intrusion detection system”. In Intelligent Systems Design and Applications (ISDA), 10th International Conference on, pages 363–367. IEEE, December 2010.

[12] Vivek K. Kshirsagar, Sonali M. Tidke and Swati Vishnu, “Intrusion Detection System using Genetic Algorithm and Data Mining: An Overview”, International Journal of Computer Science and Informatics ISSN (PRINT): 2231 – 5292, Vol-1, Iss-4, 2012.

[13] S. Mehibs and S. Hashim, “Proposed Network Intrusion Detection System Based on Fuzzy c_Mean Algorithm in Cloud Computing Environment”, JUBPAS, vol. 26, no. 2, pp. 27-35, Dec. 2017.

[14] S. Mehibs and S. Hashim, “Proposed Network Intrusion Detection System In Cloud Environment Based on Back Propagation Neural Network”, JUBPAS, vol. 26, no. 1, pp. 29-40, Dec. 2017.

[15] W. Bhaya and M. Ebadymanaa, “DDoS attack detection approach using an efficient cluster analysis in large data scale,” in 2017 Annual Conference on New Trends in Information and Communications Technology Applications, NTICT 2017, 2017.

[16] Joshi, Manish and TheyaznHassnHadi. “A Review of Network Traffic Analysis and Prediction Techniques.” CoRR abs/1507.05722 (2015): n. pag.

[17] S. Ali Khayam, F. Mirza, et al., " A SURVEY OF ANOMALY-BASED INTRUSION DETECTION SYSTEMS", School of Electrical Engineering and Computer Science, 2009.

[18] N. Moustafa, J.Slay," The Significant Features of the UNSW-NB15 and the KDD99 Data Sets for Network Intrusion Detection Systems", 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, 2015.

[19] B. Trstenjak,S. Mikac, D. Donko, "KNN with TF-IDF Based Framework for Text Categorization", 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013.

[20] P.S. Bradley, Usama Fayyad, and Cory Reina, Scaling Clustering Algorithms to Large Databases, KDD-98 Proceedings, 1998.

[21] C. Tsai, C. Lin, “A Triangle Area Based Nearest Neighbors Approach to Intrusion Detection”, Pattern Recognition, Vol. 43, No. 1, pp. 222-229, 2010.

[22] S. Mukherjee, N.Sharma, “ Intrusion Detection Using Naive Bayes Classifier with Feature Reduction”, Procedia Technology, Vol. 4, pp. 119-128, 2012.