ISSN: 2321-7782 (Online)

Volume 2, Issue 9, September 2014

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Clustering approach for Network Traffic by using Online Efficient Incremental Clustering

Shital A. Salve[1]
Computer Engineering
BSIOTR(w),wagholi
Pune, Maharashtra – India

Prof. Sanchika Bajpai[2]
Computer Engineering
BSIOTR(w),wagholi
Pune, Maharashtra – India

Abstract: Data Streams are continuous Flow of data include Network Packets, Surveillance, Financial market and day to day Business. Their volume and speed is a great challenge for data mining community to mine them A large number of researches have been proposed on intrusion detection system, which leads to the implementation of agent based intelligent IDS (IIDS), Non – intelligent IDS (NIDS), signature based IDS etc. While building such IDS models, learning algorithms from flow of network traffic plays crucial role in accuracy of IDS systems. The proposed work focuses on implementing the novel method to cluster network traffic which eliminates the limitations in existing online clustering algorithms and prove the robustness and accuracy over large stream of network traffic arriving at extremely high rate. We compare the existing algorithm with novel methods to analyse the accuracy and complexity. Online incremental efficient Clustering is adaptive; Resource aware and produce more dense and distinct cluster Than RAHc the proposed work focus on clustering and classifying real time network traffic and comparing the result of both RAHc and Online incremental efficient Clustering.

Keywords: NIDS, Data Stream Mining, streaming, RAH algorithm, Online Efficient Incremental Clustering algorithm.

I. INTRODUCTION

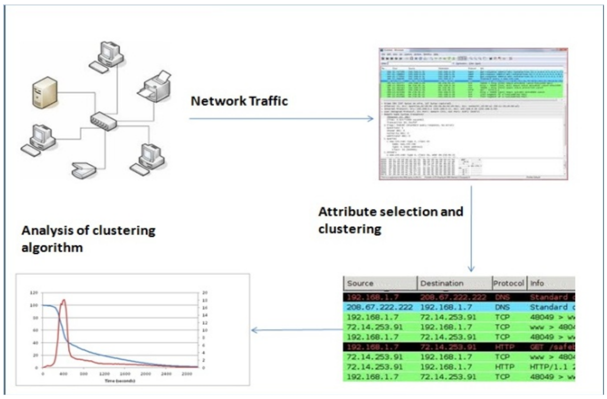
Now a days Fast growth in use of networking and internet makes security is important in recent years. The most recent topic in network security is Network Intrusion Detection System (NIDS) which keeps the security at most level. Many diverse approaches have been proposed and implemented, which minimizes the attacks and vulnerability in the network and makes it secure. Most widely used NIDS are signature based models [1]. Such models detect only known attacks, hence detecting unknown attacks without prior knowledge about specific intrusion remains a challenge. To cope with these challenges, intelligent IDS systems have evolved [2]. The IIDS system focus on specific pattern of known attacks, which reveals the root cause of intrusion by constantly learning from network traffic, and if such patterns are identified and learned, they can produce the classification model for potential intrusion. Such systems are bundled with two layers, the first layer is training or learning layer, which learns the patterns of intrusion in the flow of network traffic. Another layer is testing, which applies learned rules to detect intrusions in unknown traffic data. As learning from online data is challenging than learning from static data, it became essential to provide attention towards accuracy of stream classification algorithms [3][4].

The proposed system focus on learning from network traffic by applying innovative stream clustering algorithms and then make use of produced clusters to build classification models for IIDS. Moreover the approach justify the efficiency and simplicity of new algorithm by comparing it with existing RAH clustering algorithm.

II. IMPLEMENTATION DETAILS

The implementation is divided into three stages. The first stage is sniffing of network packets, the second stage is applying innovative online clustering algorithm and third stage is to compare existing clustering approach with new one. The system architecture shows basic components and flow of working of the system. The packet snifer is responsible for collecting the network traffic, which can be configured to filter the traffic with specific attributes. The priority attributes are then selected and arranged in the increasing order. These samples are then applied to existing RAH clustering algorithm which assigns specific cluste to every individual sample

For RAH clustering algorithm, number of clusters k , is the prior input, which is major limitation of RAH while applying it to cluster network traffic, which is diverse in nature due to which number of clusters can not be judged before clustering the data.


Network Traffic
Attribute selection and clustering
Analysis of clustering algorithm

Source	Destination	Protocol	Info
192.168.1.7	208.67.222.222	DNS	Standard
208.67.222.222	192.168.1.7	DNS	Standard
192.168.1.7	72.14.253.91	TCP	48049 > w
72.14.253.91	192.168.1.7	TCP	www > 4804
192.168.1.7	72.14.253.91	TCP	48049 > w
192.168.1.7	72.14.253.91	HTTP	GET /safe
72.14.253.91	192.168.1.7	TCP	www > 4804
72.14.253.91	192.168.1.7	HTTP	HTTP/1.1
192.168.1.7	72.14.253.91	TCP	48049 > w

Fig. 1. Architecture of proposed system

Next, the same samples are applied to Online Efficient Incremental Clustering algorithm, where number of clusters, k is not the prior input .The clusters are fromed according to diverse nature of traffic data. At the end we are comparing the results for accuracy and complexity of both the algorithms and justify that Online Efficient Incremental Clustering is best suitable approach to cluster the netwok traffic. The basic component of system are packet sniffer, attribute and their priorities and clustering algorithms.

1. Network traffic and Packet Sniffer

The packet sniffer sniffs the incoming packets through the network adapter. The sniffer is designed such that user can configure the attributes of packet that are traced by the sniffer. Many studies have revealed that the attributes such as Source IP Address, Destination IP Address, Source Port Number, Destination Port Number, TCP Window Size, and TCP Data Length are most promising fields associated with different types of attacks [10], hence the above fields are considered while applying clustering algorithm on the stream of packets.

2. Attribute and their Priority Selection

Input to the clustering algorithm is mostly one or two dimensional numeric data points. By having single dimensional data, the similarity between two samples can be computed by taking direct difference between two values. For two dimensional sample data, the similarity between two samples is computed by Euclidian or Manhattan distance measures. But for multi-dimensional categorical data, the difference measure is very challenging. To calculate distance among network packets there is no standard measure. The packet is set of attributes and every attribute may have numeric or categorical values. To compute the similarity among packets, first we need to focus on specific attribute values. If such selected attribute values for both packets are equal, then we can state that two packets are similar. But predicting such similarity on the basis of single attribute would not give accuracy, so we require multiple attributes and their priorities. The input to algorithm is two packet samples and three attributes with increasing priorities. The algorithm compares every attribute of first sample with every attribute of second sample, if both attributes are equal, then it checks their priorities from available priority attributes. If the priority is highest, it increases the weight by weight factor 4, if priority is normal, then by 3 and if the priority is low then by 2. Finally it returns the weight.

A. Online Clustering

1) Online Efficient Incremental Clustering

Predefined number of cluster (k) and threshold values prior to online clustering are bottlenecks for online learner. The innovative Online Efficient Incremental Clustering algorithm copes with above bottlenecks and proves its ability to learn online without having predefined number of clusters (k) and any threshold values.

The algorithm initializes data – centre threshold (DC_{TH}) and centre – centre threshold (CC_{TH}) values as 0. It then read first available sample S. If S is the only sample in cluster space then the sample S would be the first member of new cluster. If S is not the only sample, then algorithm computes distance of sample S with centre of all available clusters. It then computes minimum distance Di. Suppose the index of cluster to which S is having minimum distance is p.By comparing sample S with all available centres, it yields three possible scenarios. In first scenario, the distance between sample S and cluster centre C[p] is 0. In this case sample S is merged into the cluster C[p]. The cluster centre of C[p] is updated and CC_{TH} also updated as minimum of available $CC_{TH.}$ In second scenario the distance between sample S and cluster centre C[p] is greater than $CC_{TH.}$ In this case the new cluster is formed having S as the member of that cluster. After that CC_{TH} is updated. In third scenario, the distance between sample S and C[p] is less than CC_{TH}. In this case sample S is merged into cluster C[p], CC_{TH} and DC_{TH} are updated. If DC_{TH} is greater than the CC_{TH}, then cluster C[p] is spitted to satisfy $CC_{TH} > DC_{TH}$.

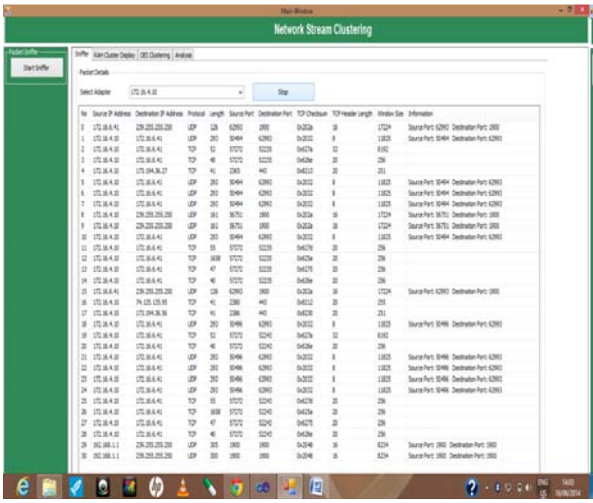
III. Comparison Attributes

We have compared RAHc algorithm and online efficient incremental algorithm and proving that novel algorithm is best by using following comparison attributes

1. SSQ, The quality of the clustering is measured by the sum of squared distances (SSQ) of data points from their assigned me- dians. The goal is to _and a set of k medians which minimize the SSQ measure. The generalized optimization problem, in which any distance metric substitutes for the squared Euclidean distance, is known as the k{Median problem).
2. Intercluster distance, the Inter-cluster distance measured by within-cluster sum of squares. Measures cluster "compactness”, when intercluster distance is max then it should be a good cluster.
3. Intracluster distance this distance we compute the finding out distance between two cluster

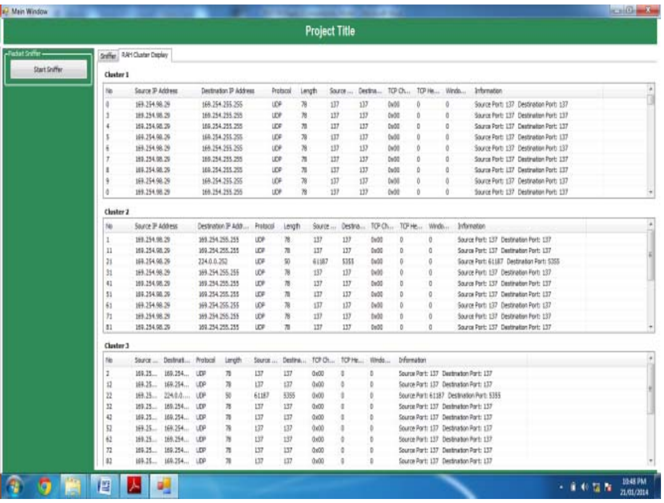
IV. Results and Data Set

The implemented approach shows the network traffic captured by packet sniffer. The packet sniffer is configured to capture packets with attributes – source port, destination port, source IP address, destination IP address, TCP length, TCP checksum. In second window, the module clusters every network packet into specific category of cluster using RAH algorithm. For calculating similarity measurement among packet, three attributes are selected in their incremental priority order. These three attributes are source IP address, destination port number and TCP header length.



No	Source IP Address	Destination IP Address	Protocol	Length	Source Port	Destination Port	TCP Checksum	TCP Header Length	Window Size	Information

Fig. 2. Network traffic captured by packet sniffer



Cluster 1

No	Source IP Address	Destination IP Address	Protocol	Length	Source ...	Destina...	TCP Ch...	TCP He...	Windo...	Information

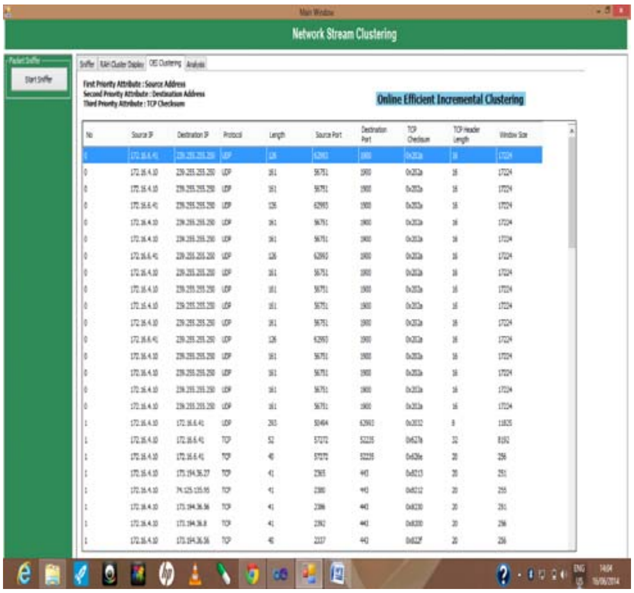
Cluster 2

No	Source IP Address	Destination IP Addr...	Protocol	Length	Source ...	Destina...	TCP Ch...	TCP He...	Windo...	Information

Cluster 3

No	Source ...	Destinati...	Protocol	Length	Source ...	Destina...	TCP Ch...	TCP He...	Windo...	Information

Fig. 3. Three clusters of packets using RAH clustering algorithm.

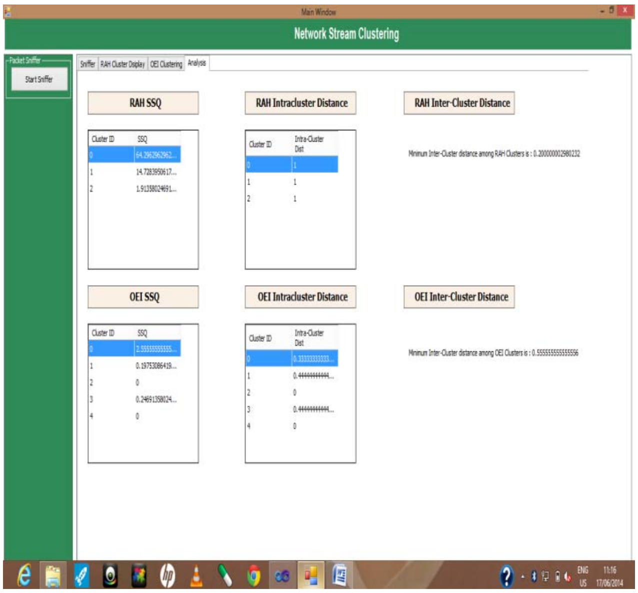


First Priority Attribute : Source Address
Second Priority Attribute : Destination Address
Third Priority Attribute : TCP Checksum

Online Efficient Incremental Clustering

No	Source IP	Destination IP	Protocol	Length	Source Port	Destination Port	TCP Checksum	TCP Header Length	Window Size

Fig 4.output of online efficient incremental clustering


Main Window
Network Stream Clustering
Packet Sniffer
Start Sniffer
Sniffer
RAH Cluster Display
OEI Clustering
Analysis
RAH SSQ
RAH Intracluster Distance
RAH Inter-Cluster Distance
OEI SSQ
OEI Intracluster Distance
OEI Inter-Cluster Distance

Cluster ID	SSQ
0	64.2962962962...
1	14.7283950617...
2	1.91358024691...

Cluster ID	Intra-Cluster Dist
0	1
1	1
2	1

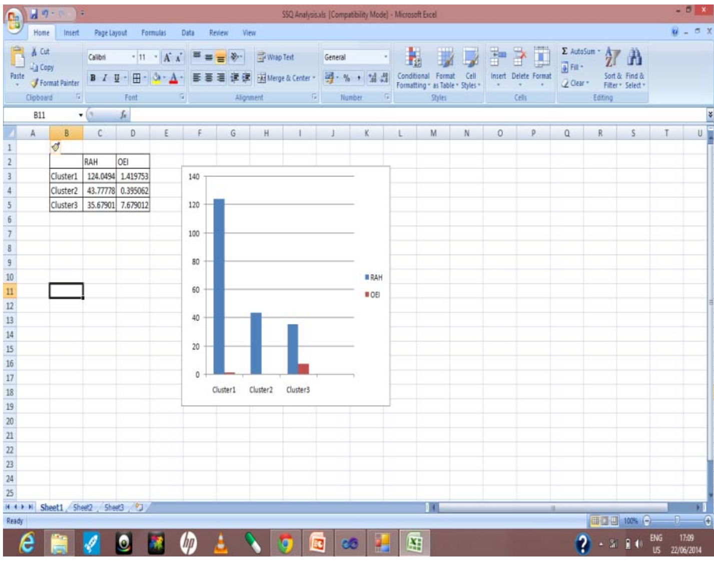
Minimum Inter-Cluster distance among RAH Clusters is : 0.200000002980232

Cluster ID	SSQ
0	2.55555555555...
1	0.19753086419...
2	0
3	0.24691358024...
4	0

Cluster ID	Intra-Cluster Dist
0	0.33333333333...
1	0.44444444444...
2	0
3	0.44444444444...
4	0

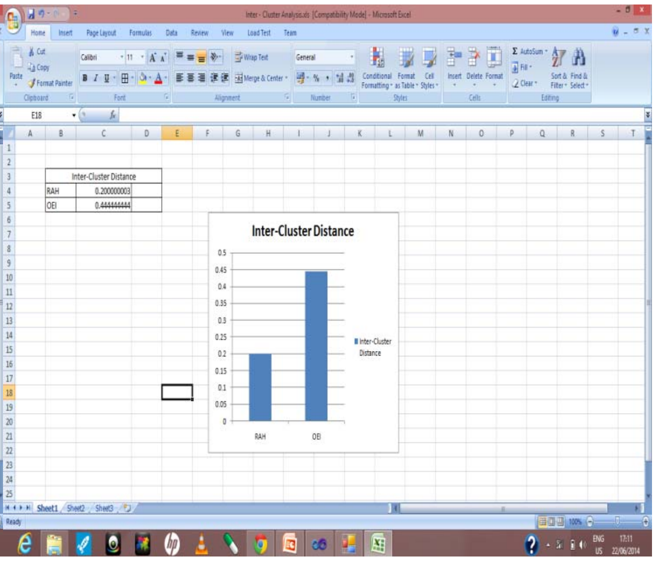
Minimum Inter-Cluster distance among OEI Clusters is : 0.555555555555556

fig 5. comparison between RAHc and OEI



	RAH	OEI
Cluster1	124.0494	1.419753
Cluster2	43.77778	0.395062
Cluster3	35.67901	7.679012

Fig 6.result of SSQ compared with RAHc and OEI



Inter-Cluster Distance	
RAH	0.200000003
OEI	0.444444444

Fig 7. Result of intercluster Distance With RAHc and OEI

V. CONCLUSION AND FUTURE ENHANCEMENT

The Online Efficient Incremental Clustering proves its effectiveness as it does not requires the predefined number of cluster (k) and threshold values prior to the clustering. For clustering network data with extremely high rate, the above values are mostly unknown. And if stated would produce wrong results. The algorithm is best suited for such online clustering with high accuracy. Moreover the time complexity of resource aware learner is high due to the threshold bound checking and convergences of algorithm, which is completely eliminated in Online Efficient Incremental Clustering hence it is less complex than existing approaches. Partially formed clusters can be applied classification models to classify the data, hence applicable to IDS system. The clustering algorithms can be applied to cluster the network traffic in wireless LAN. The model can be applied to proxy system, where it is easy to analyze the most visited IP's and various type of application on the basis of clusters of packets formed by selecting specific attributes and their priorities.

References

1. FarooqAnjum, DhanantSubhadrabandhu and SaswatiSarkar,‖ ―Signature based Intrusion Detection for Wireless Ad-Hoc Networks: A Comparative study of various routing protocols‖, Telcordia. Tech Inc. Morristown NJ 07960J.
2. N.Jaisankar and R.Saravanan K. DuraiSwamy,‖Intelligent Intrusion Detection System Framework Using Mobile Agents‖, International Journal of Network Security & Its Applications (IJNSA), Vol 1, No 2, July 2009R.
3. Pedro Domingos, Geoff Hulten, ―Mining High Speed Data Streams‖.
4. Charu C. Aggarwal, Jiawei Han, Jianyong Wang, Philip S. Yu,‖ A Framework for Clustering Evolving Data Streams‖, Proeedings of the 29th VLDB Conference, Berlin, Germany, 2003.
5. Andreas Kind, Marc Ph. Stoecklin, and Xenofontas Dimitropoulos,‖ Histogram-Based Traffic Anomaly Detection‖, IEEE Transactions On Network Service Management, Vol. 6, No. 2, June 2009
6. Jiankun Hu and Xinghuo Yu,‖ A Simple and Efficient Hidden Markov Model Scheme for Host-Based Anomaly Intrusion Detection‖
7. Jake Ryan, Meng-Jang Lin, ―Intrusion Detection with Neural Networks‖, Advances in Neural Information Processing Systems 10, Cambridge,MA:MIT Press, 1998.
8. Vivek K. Kshirsagar, Sonali M. Tidke& 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.
9. Ching-Ming Chao and Guan-Lin Chao,‖ Resource-Aware High Quality Clustering in Ubiquitous Data Streams‖, in Proceedings of the 13th International Conference on Enterprise Information Systems, Beijing, China (2011).
10. Anna Sperotto, GregorSchaffrath, RaminSadre, CristianMorariu, AikoPras and Burkhard Stiller,‖ An Overview of IP Flow-Based Intrusion Detection‖, IEEE Communications Surveys & Tutorials, Vol. 12, No. 3, Third Quarter 2010