K-Means Clustering Implementation in Network Traffic Activities

Purnawansyah
Department of Informatics Engineering
Faculty of Computer Science
Universitas Muslim Indonesia - Indonesia
purnawansyah@gmail.com

Haviluddin
Department of Computer Science
Faculty of Computer Science and Information Technology
Mulawarman University - Indonesia
haviluddin@gmail.com

***Abstract*—At present, management analysis bandwidth in a university is indispensable. It aims to control bandwidth usage, so that all spots can be served comfortably especially to supporting the teaching and learning activities. In this study, an analysis and clustering of the university internet traffic is required as bandwidth management decision support. Therefore, K-Means as a clustering algorithm bandwidth usage was implemented and explored. The results showed that the K-Means method can perform clustering with 3 and 4 clusters. The cluster is described high, medium and low bandwidth usage at certain times of each unit. Furthermore, the clustering result could be a recommendation management bandwidth for network administrator in order to planning, sharing, and controlling bandwidth.**

Keywords—clustering; K-Means; university; network traffic

I. INTRODUCTION

Management of bandwidth capacity is indispensable. In order to reduce the waste of bandwidth usage, steady access, helps network administrators to control bandwidth usage as well as. Although, network administration system is able to capture the use of bandwidth (i.e., Mikrotik, Wireshark, Splunk, OpManager, CACTI, etc.). However, these tools in analyzing bandwidth usage are still less capable.

In order to overcome the weaknesses of these tools, especially in terms of data analyzing. Later, a method of data analysis bandwidth management is required. Furthermore, the mapping (cluster) will provide a model bandwidth distribution to support the network administrator's decision especially in providing the bandwidth at any point or spot.

In this paper, cluster network traffic by using machine learning methods. Then, K-Means as a clustering method was implemented. The purpose of this study was to test the feasibility of K-Means clustering and how it works in the real world problems, especially on network traffic. The rest of this paper is organized as following; recent related work is discussed and presented in section 2. In section 3, is proposed the K-Means clustering approach. In section 4 a set of experiments is presented K-Means clustering methods. Finally, conclusion and future work.

II. RELATED WORK

In the recent years, network traffic clustering has been introduced as an efficient method for managing and operating bandwidth distributions collections by network administrators in response to clients' queries. Many clustering techniques are implemented. For example, using K-Means has proved to be an effective clustering method widely used in various fields such as hydrology [1, 2], pattern recognition [3], medical image [4, 5], network [6, 7], engineering [8]. In managing internet traffic [9] have been used K-Means to analyzed user access behavior campus network at the Nantong University, China. The dataset was used 2000 student users. The results showed that clustering played important rule in order to analyzed different behavior of the campus network. Then, the K-Means was capable as a model to clustering the datasets. Later, [7] have been used K-Means, Spectral Filtering, Newman's, Fast SVD, and Genetic Algorithm (GA) to cluster 220 member of Abilene (now known as Internet2 Network) is the U.S. The result indicated that combined method outperforms a good clusters than one method. Then, [10] have been compared LEACH-Centralized, KMeans-CP, FCM-CP and HSA-CP protocols to clustering and data delivery process for various realistic topologies. The simulations were carried out for 50 nodes to 200 nodes, with different BS locations. The results proposed that HSACP was report good amount of data, if applied to unknown or strange deployment conditions.

III. METHODOLOGY

In this section, a brief information on the general network traffic clustering model will be presented by using K-Means model.

A. Clustering

Clustering is a method of analyzing data. The aim is to group data with similar characteristics to a 'same-area' and the data with different characteristics to the 'different-area' [8, 11]. Nevertheless, an essential excellent issue is the dissimilarity measurement data or objects in the group. In this study, the Euclidean Distance method for measuring the shortest distance between two data has been used.

978-1-5090-5548-7/16/$31.00 ©2016 IEEE

B. *Principle of K-Means*

K-Means is one of many methods used to perform clustering that included in a group of unsupervised methods. K-Means method introduced gradually by some researchers as Steinhaus in 1956, Lloyd 1957, Forgy/Jancey in 1965/1966, and James MacQueen's sequential k-means algorithm in 1965/1967 [12, 13]. K-Means clustering algorithm is one of the best-known and most well-known clustering algorithms utilized in a variety of areas. In principle, K-Means clustering algorithm works on the assumption that the initial centers are given. The search for the final clusters or centers starts from these initial centers. K points from the dataset as the initial cluster center, putting the sample to the class where the nearest cluster center in. Then, the distances of all data elements are calculated by distances formula, e.g., Euclidean distance, Manhattan, Cosine, Chebychev, Minkowski, Tanimoto, etc [5, 14]. The K-Means flowchart can be seen in Fig. 1.

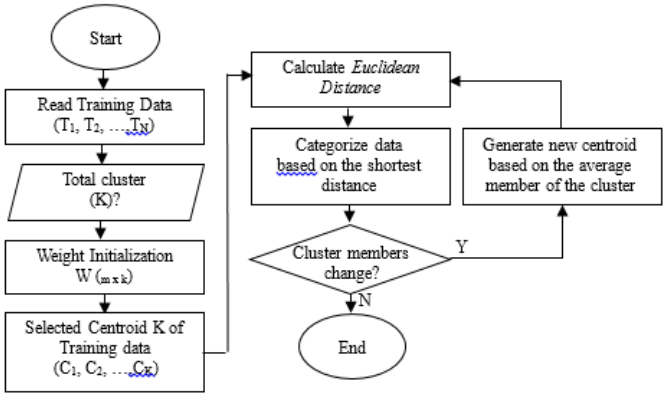

Fig. 1. K-Means Flowchart

In this study, the K-Means algorithm steps are as follow:

1. Read training data (T_1, T_2, .., T_n), where T_n is data training
2. Choose a number of desired clusters, *k*.
3. Choose *k* starting points to be used as initial estimates of the cluster centroids. These are the initial starting values.
4. Calculate Distance using some distance method (i.e., Euclidean Distance, Manhattan, Cosine, Chebychev, Minkowski, Tanimoto etc.) for each training data.
5. Examine each point in the given dataset and assign it to the cluster whose centroid is nearest to it.
6. When each point is assigned to a cluster, recalculate the new *k* centroids.
7. Repeat steps 3 and 4 until no point changes its cluster assignment, or until a maximum number of passes through the data set is performed.

Then, the Pseudocode of the K-Means algorithm as follow:

Place randomly the *K* cluster centers

While not stop criterion ***do***

For each object ***do***

Compute distance measure to each cluster

Assign it to the closest cluster

End for

Recalculate the cluster centers positions (2)

End while

C. *Data Collection and Data Transformation*

In this paper, the datasets recorded from ICT unit of Universitas Mulawarman for 5 months (from January to May 2016) or 456 datasets were collected and explored. In order to achieve the clustering results is pretty good, so a normalization data process is required [15, 16]. In this paper, the datasets before going through the network are normalized between 0.05 and 1. If X, X_{max}, X_{min} are the original, maximum and minimum values of the raw data, respectively then the normalization of X called X_n, can be obtained by the following transformation function, (1).

$$X_n = 0.05 + 0.95 \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

The plot of network traffic dataset after normalization can be seen in Fig. 2.

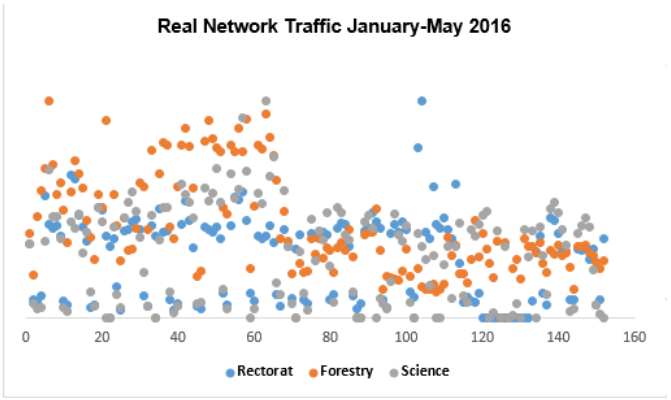

Fig. 2. Real Network Traffic January-May 2016

D. *Performance Metrics*

Euclidean Distance

Euclidean Distance is a metric used to calculate the similarity of two vectors [14]. Euclidean Distance formula is the square root of differences between two vectors. In principle, the Euclidean distance is to calculate the distance in the "distance-space" by calculating two points in the shortest distance. In this study, Euclidean distance was used to calculate the distance K-Means algorithm and haven't combined with other distance methods. In this study, Euclidean distance was implemented in order to measure the performance clustering model. The Euclidean measure corresponds to the shortest geometric distance between two points [16]. The formulation as shown below, (2).

$$d = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \quad (2)$$

Where, d is distance between the object; N is data dimension; x_i and y_i are object coordinate in dimensions.

IV. Experimental and Discussion Results

This section presents the results obtained for the K-Means algorithm in clustering network traffic datasets. In the K-Means algorithm, based on the determination of the distance between the centroid of the training data. The number is based on the cluster centroid desired. Then, the centroid number is randomly generated by considering the min-max training data. In each iteration, calculated the distance of each training data with centroids. Then, correcting new centroid value is based on the average value of each cluster members. Then, an iteration is stopped, if the cluster members have not changed. In this study, MATLAB R2013b as a tool for clustering was utilized.

In this training scheme, training data will be grouped 3 and 4 clusters. The aim is to observe good grouping patterns. The average value of data as one of the centroid value is used. Later, another centroid value is determined randomly. The training results are shown in Fig. 2.

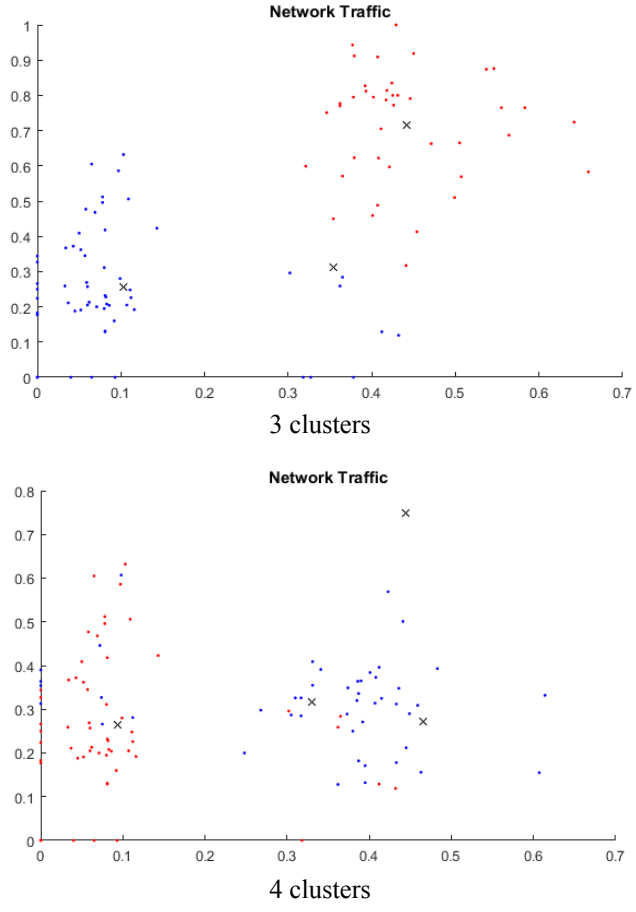

Fig. 3. K-Means Algorithm Results

TABLE I. Network Traffic Cluster Results

Cluster	Pattern	Rectorat	Forestry	Science	Cluster	Pattern	Rectorat	Forestry	Science
3	High	0.433	0.467	0.109	3	High	0.648	0.464	0.274
2	Med	0.863	0.116	0.442	3	High	0.580	0.431	0.087
2	Med	0.731	0.255	0.468	3	High	0.503	0.530	0.120
2	Med	0.798	0.339	0.560	2	Med	0.808	0.198	0.403
1	Low	0.272	0.751	0.441	2	Med	0.904	0.061	0.464
1	Low	0.328	1.096	0.785	3	High	0.431	0.611	0.169
1	Low	0.278	0.702	0.408	3	High	0.448	0.580	0.143
3	High	0.354	0.617	0.286	3	High	0.489	0.510	0.084
1	Low	0.213	0.662	0.361	3	High	0.490	0.422	0.118
2	Med	0.737	0.259	0.478	2	Med	0.869	0.096	0.451
2	Med	0.897	0.128	0.525	2	Med	1.108	0.281	0.637
1	Low	0.293	0.820	0.438	2	Med	1.100	0.276	0.625
1	Low	0.272	0.853	0.518	3	High	0.403	0.553	0.121
1	Low	0.081	0.839	0.482	3	High	0.384	0.586	0.154
1	Low	0.272	0.642	0.356	3	High	0.423	0.572	0.158
1	Low	0.393	0.831	0.525	3	High	0.644	0.449	0.430
2	Med	0.738	0.195	0.410	3	High	0.617	0.432	0.310
2	Med	0.784	0.240	0.440	2	Med	0.978	0.144	0.519
1	Low	0.176	0.777	0.380	2	Med	0.918	0.100	0.495
1	Low	0.243	0.808	0.431	3	High	0.725	0.333	0.256
1	Low	0.581	0.783	0.719	3	High	0.601	0.566	0.187
2	Med	0.902	0.490	0.514	3	High	0.663	0.452	0.169
1	Low	0.240	0.626	0.290	3	High	0.627	0.464	0.153
2	Med	0.766	0.181	0.421	3	High	0.520	0.495	0.089
2	Med	0.926	0.087	0.520	2	Med	0.923	0.091	0.475
1	Low	0.268	0.698	0.314	2	Med	1.087	0.272	0.610
3	High	0.411	0.761	0.365	3	High	0.644	0.798	0.444
1	Low	0.453	0.906	0.523	3	High	0.844	0.978	0.674
1	Low	0.336	0.825	0.443	3	High	0.727	0.392	0.228
1	Low	0.200	0.682	0.368	3	High	0.755	0.350	0.294
3	High	0.522	0.455	0.429	3	High	0.720	0.570	0.321
2	Med	0.836	0.166	0.501	2	Med	0.954	0.369	0.502
1	Low	0.479	1.178	0.875	2	Med	0.936	0.344	0.484
2	Med	0.909	0.466	0.495	3	High	0.706	0.434	0.238
1	Low	0.223	0.866	0.546	3	High	0.497	0.447	0.100
1	Low	0.287	0.792	0.519	2	Med	0.816	0.275	0.393
1	Low	0.135	0.837	0.529	3	High	0.572	0.593	0.274
2	Med	0.797	0.179	0.456	3	High	0.654	0.421	0.185
2	Med	0.753	0.329	0.492	2	Med	0.909	0.082	0.470
2	Med	0.727	0.358	0.536	2	Med	0.952	0.116	0.499
1	Low	0.197	1.037	0.690	3	High	0.665	0.373	0.287
1	Low	0.193	1.030	0.689	3	High	0.508	0.468	0.328
1	Low	0.097	0.941	0.606	3	High	0.640	0.388	0.256
1	Low	0.195	0.682	0.356	3	High	0.617	0.475	0.375
2	Med	0.930	0.085	0.497	3	High	0.611	0.550	0.398
2	Med	0.845	0.178	0.454	2	Med	0.981	0.141	0.580
1	Low	0.198	1.033	0.694	2	Med	1.004	0.152	0.578
1	Low	0.203	0.950	0.656	3	High	0.661	0.393	0.361
1	Low	0.156	0.979	0.654	2	Med	1.081	0.283	0.625
1	Low	0.181	0.975	0.619	2	Med	0.708	0.312	0.371
1	Low	0.252	1.009	0.693	2	Med	1.090	0.282	0.631
2	Med	0.766	0.270	0.475	2	Med	0.986	0.138	0.575
2	Med	0.837	0.235	0.536	2	Med	0.881	0.113	0.494
1	Low	0.231	1.068	0.714	2	Med	1.012	0.156	0.588
1	Low	0.129	0.937	0.575	3	High	0.599	0.475	0.373
1	Low	0.281	0.962	0.637	2	Med	0.714	0.282	0.371
1	Low	0.572	1.388	0.989	3	High	0.614	0.387	0.282
1	Low	0.253	1.066	0.731	2	Med	0.852	0.219	0.447
2	Med	0.951	0.105	0.526	3	High	0.486	0.483	0.067
2	Med	0.832	0.271	0.547	2	Med	0.809	0.233	0.369
1	Low	0.183	0.852	0.538	2	Med	0.924	0.103	0.486
1	Low	0.100	0.870	0.556	3	High	0.439	0.637	0.194
1	Low	0.575	1.215	0.890	3	High	0.536	0.598	0.197
1	Low	0.183	0.972	0.632	3	High	0.447	0.562	0.125
1	Low	0.245	1.009	0.653	3	High	0.480	0.521	0.160
2	Med	0.723	0.394	0.525	3	High	0.491	0.482	0.169
2	Med	0.843	0.139	0.475	2	Med	0.916	0.053	0.495
1	Low	0.267	0.728	0.315	2	Med	0.900	0.185	0.471
3	High	0.478	0.433	0.079	3	High	0.453	0.505	0.133
2	Med	0.938	0.094	0.499	3	High	0.450	0.518	0.132
2	Med	1.040	0.348	0.555	3	High	0.414	0.593	0.207
3	High	0.740	0.376	0.281	3	High	0.502	0.473	0.123
2	Med	0.915	0.105	0.473	3	High	0.614	0.313	0.163
2	Med	0.958	0.101	0.530	2	Med	0.895	0.049	0.485

TABLE I. NETWORK TRAFFIC CLUSTER RESULTS (*CONTINUE*)

Cluster	Pattern	Rectorat	Forestry	Science	Cluster	Pattern	Rectorat	Forestry	Science
3	High	0.640	0.503	0.283	2	Med	0.938	0.077	0.516
3	High	0.731	0.347	0.289	2	Med	0.762	0.308	0.379

In this experiment, the purpose of the cluster network traffic is to investigate bandwidth users in the pattern (high, medium, and low) that over a period of 30 minutes per days in five months from three units. Based on the experiments, it appears that grouping 3 and 4 clusters have been produced the same number of clusters; three clusters. In other word, the cluster results was consistent.

Based on Table I, the high usage of network traffic from the third units at Universitas Mulawarman was from beginning until mid-months. Then, the visual of clustering can be seen in Fig. 3.

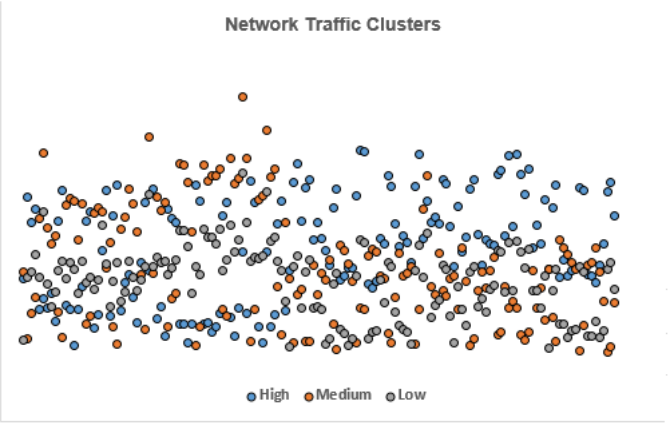

Fig. 4. High-Medium-Low K-Means Clustering Results

V. CONCLUSIONS

In this study, network traffic data of Universitas Mulawarman ICT Unit was analyzed. The K-Means algorithm was applied to identify bandwidth user's pattern (high, medium, and low) per day over five months of each unit. The output clusters were evaluated using Euclidean distance as a performance of the proposed method. The results showed that K-Means method with Euclidean distance as a performance metrics was able to produce a good cluster, especially in network traffic clustering. In this study, the pretty good bandwidth usage clustering results were 3 and 4 clusters.

REFERENCES

[1] Z. Zahmatkesh, M. Karamouz, and S. Nazif, "Uncertainty based modeling of rainfall-runoff: Combined differential evolution adaptive Metropolis (DREAM) and K-means clustering," Advances in Water Resources, vol. 83 (2015), pp. 405–420, 2015.

[2] S. M. Sharif, F. M. Kusin, Z. H. Asha'ari, and A. Z. Aris, "Characterization of water quality conditions in the Klang River Basin, Malaysia using self organizing map and K-means algorithm," in International Conference on Environmental Forensics 2015 (iENFORCE2015), Malaysia., 2015, pp. 73 – 78.

[3] S. Soheily-Khah, A. Douzal-Chouakria, and E. Gaussier, "Generalized k-means-based clustering for temporal data under weighted and kernel time warp," Pattern Recognition Letters, vol. 75 (2016), pp. 63–69, 2016.

[4] R. HariKumar, B. Vinothkumar, and G. Karthick, "Performance Analysis for Quality Measures Using K means Clustering and EM Models in Segmentation of Medical Images," International Journal of Soft Computing and Engineering (IJSCE), vol. 1, Issue-6, January 2012, pp. 74-80, 2012.

[5] Y. G. Jung, M. S. Kang, and J. Heo, "Clustering performance comparison using K-means and expectation maximization algorithms," Biotechnology & Biotechnological Equipment, vol. Vol. 28, No. S1, pp. 44-48, 2014.

[6] M. G. Khair, B. Kantarci, and H. T. Mouftah, "Heterogeneous Clustering Of Sensor Network," Procedia Computer Science, vol. 5 (2011), pp. 939–944, 2011.

[7] M. Naldi, S. Salcedo-Sanz, L. Carro-Calvo, L. Laura, A. Portilla-Figueras, and G. F. Italiano, "A traffic-based evolutionary algorithm for network clustering," Applied Soft Computing, vol. 13 (2013), pp. 4303–4319, 2013.

[8] K. Lee, S. Jung, T. Lee, and J. Choe, "Use of Clustered Covariance and Selective Measurement Data in Ensemble Smoother for Three-Dimensional Reservoir Characterization," J. Energy Resour. Technol, vol. 2016;139(2), 2017.

[9] Q. Shi, L. Xu, Z. Shi, Y. Chen, and Y. Shao, "Analysis and Research of the Campus Network User's Behavior based on K-means Clustering Algorithm," in 2013 Fourth International Conference on Digital Manufacturing & Automation, 2013, pp. 196-201.

[10] G. Raval, M. Bhavsar, and N. Patel, "Analyzing the Performance of Centralized Clustering Techniques for Realistic Wireless Sensor Network Topologies," in 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015), 2015, pp. 1026 – 1035.

[11] V. Sucasas, A. Radwan, H. Marques, J. Rodriguez, S. Vahid, and R. Tafazolli, "A survey on clustering techniques for cooperative wireless networks," Ad Hoc Networks, vol. 47 (2016), pp. 53–81, 2016.

[12] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, 1967, pp. 281–297.

[13] Y. S. Thakare and S. B. Bagal, "Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics," International Journal of Computer Applications, vol. 110 – No. 11, January 2015, pp. 12-15, 2015.

[14] A. Singh, A. Yadav, and A. Rana, "K-means with Three different Distance Metrics," International Journal of Computer Applications, vol. 67– No.10, April 2013, pp. 13-17, 2013.

[15] D. Ballabio and M. Vasighi, "A MATLAB toolbox for Self Organizing Maps and supervised neural network learning strategies," Chemometrics and Intelligent Laboratory Systems, vol. 118 (2012), pp. 24–32, 2012.

[16] T. Kohonen, "Essentials of the self-organizing map," Neural Networks, vol. 37, pp. pp. 52–65, 2013.