Advanced Science Letters
Vol. 4, 400–407, 2015

A Performance Neighborhood Distance (*ndist*) between K-Means and SOM Algorithms

Mislan[1], Haviluddin[2], Rayner Alfred[3], Achmad Fanany Onnilita Gaffar[4]

[1]Faculty of Mathematics and Natural Science, Universitas Mulawarman, Indonesia
[2]Faculty of Computer Science and Information Technology, Universitas Mulawarman, Indonesia
[3]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia
[4]Dept. of Information Technology, State Polytechnic of Samarinda, Indonesia

Clustering is an important mean of data mining based on separating data categories based on similar features. This paper aims to compare the performance of neighborhood distance (*ndist*) between *k-Means* and Self-Organizing Maps (SOM) algorithms. The sample datasets used in this study include rainfall datasets obtained from thirteen stations located in East Kalimantan. This paper outlines and presents the comparison of performance of the *ndist* for both the *k-Means* and SOM in analyzing and clustering rainfall datasets. The performances of these algorithms are compared based on the *ndist* values. The findings of this study indicated that the *k-Means* has been proved to be effective in terms of the performance of the *ndist* by using the centroid concept better than SOM algorithm. This paper is concluded by recommending some future works that can be applied in order to improve the *ndist* of the K-Means and SOM.

Keywords: Clustering, K-Means, Self-Organizing Maps, *ndist*, Rainfall.

1. INTRODUCTION

Clustering is one of the most extensively used techniques for exploratory data analysis of data mining. Clustering belongs to the group of unsupervised algorithms. Clustering is a widely discussed issue which is found in many domains. Researchers from various disciplines (finance, mathematics, physics, biology, etc.) have addressed it. Trying to categorize the data or finding the groupings among the data is not a simple task for humans unless the data is low dimensionality. This is the reason some methods in machine learning have been proposed to solve these kinds of problems. The idea of data clustering is usually tend to summarize this large amount of data into a small number of groups or categories in order to further facilitate its analysis[5, 10].

*Email Address: haviluddin@unmul.ac.id

Therefore, in the literature, several researchers have investigated the application of clustering, classification, and prediction tasks in analyzing datasets obtained from the hydrology field[4, 8, 13]. Many researchers recommended that machine learning approaches are highly recommended to extract valuable information from the data. The *k*-Means and SOM algorithms have been used to determine the spatiotemporal pattern of the characteristics water quality data and identifying the sources of pollution in the Klang River Basin, Malaysia. The results proposed that *k*-Means and SOM can be useful in clustering these datasets based on water quality[10].

In order to address the issue in flood hazard analysis, [14] differential evolution adaptive Metropolis (DREAM) and *k*-Means clustering will be implemented. The historical daily data for precipitation, temperature and stream flow are obtained from National Oceanic and Atmospheric Administration (NOAA)[15] and U.S. Geological Survey (USGS)[16].

 1936-6612/2011/4/400/008 doi:10.1166/asl.2011.1261

The experiment reveals that the DREAM and *k*-Means combination was successfully applied and effective for watershed characteristics, and also these algorithms can be used in identification of the main factors affecting flood hazard analysis. [9]SOM combined with Gibbs Diagrams to investigate the seasonal and spatial hydro-geochemical characteristics of groundwater in the Pleistocene confined aquifer of the Red River Delta, Vietnam. The results indicated that SOM combine Gibbs Diagrams was very effective tool for the assessment of groundwater quality in terms of the seasonal and spatial hydro-geochemical characteristics.

In this paper, two representative clustering algorithms are implemented and discussed, namely *k-Means* and Self-Organizing Maps (SOM) clustering algorithms. The efficiency of these clustering algorithms depend on how efficiently the each data point satisfies the common properties of its cluster and how much dissimilar these data points are to the other clusters. In this study, for comparing the performance of these algorithms, neighborhood distance (*ndist*) is used to describe measure for a good cluster. The remainder of this paper is structured as follows. Section 2 summarizes the literature related to forecast models. In Section 3, three algorithms techniques are applied to predict network traffic datasets with the technique performance discussed. Finally, conclusions and recommendations for future work are provided in Section 4.

2. METHODOLOGY

Clustering can be used to produce descriptive analysis for any datasets. Therefore, a cluster is a collection of objects which are 'similar' between each other and are 'dissimilar' to the objects belonging to other clusters[1]. In this research, two of clustering algorithms namely *k*-Means and SOM have been explored and compared by using rainfall datasets. In this section, brief principles of *k*-Means and SOM techniques for clustering are briefly described.

2.1. THE PRINCIPLE OF K-MEANS

The *k*-Means clustering algorithm is one of the best-known and most popular clustering algorithms used in a variety of domains. In principle, *k*-Means clustering algorithm works on the assumption that the initial centers are provided. 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 (ED), Manhattan, Cosine, Chebychev, Minkowski, Tanimoto, etc [5, 11, 12]). This process is illustrated in Fig. 1.

The cluster analysis procedure is analyzed to determine the properties of the dataset and the target variable. It is typically used to determine how to measure similarity distance [3]. In this study, the *k*-Means algorithm steps are as follow:

1. Read data (T_1, T_2, .., T_n)
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 Euclidean Distance each 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.

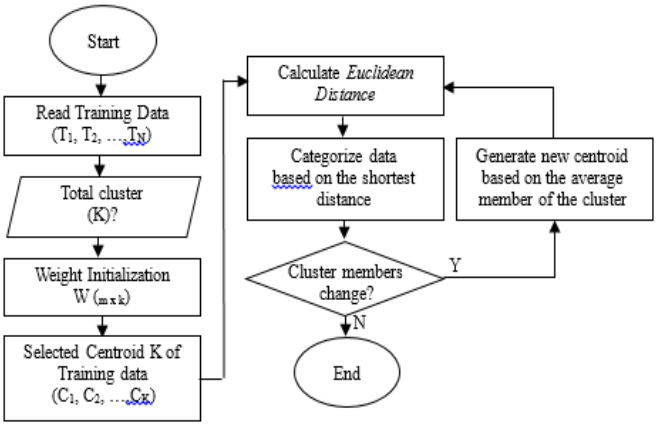


Fig.1. K-Means Flowchart

2.2. THE PRINCIPLE OF SELF-ORGANIZING MAPS (SOM)

Self-Organizing Maps (SOM) is classified as an unsupervised learning, which is one type of neural network (NN) proposed by Prof. Teuvo Kohonen[2,7]. The SOM algorithm is considered as one of powerful tools for describing hidden information in large datasets. The SOM architecture consists of *n* vectors training on input layer, *m* unit category or cluster on output layer, intra-layer connecting between the inputs to outputs layers, Fig. 2. Each neuron in the input layer is directly connected, where each relationship has a weighting vector of length *n*. The intra-layer used in the weight update algorithm is *m x n* matrix, where *m* is the number of desired clusters. In principle, Intra-layer is a function of the neighborhood distance between data in the input layer. This function will map the data into the membership of each cluster in the output layer [6, 7]. The SOM flowchart is shown in Fig. 3. The neighborhood distance as shown in Equation below.

$$L^2 = \sum_{s=1}^{k} (P_{i=1,s} - W_{j=1,s})^2 \quad (1)$$

where, L^2 is Euclidean distance; P is data input; n is amount of data; W is intra-layer weight; $i = 1$ is data vector to cluster $j = 1$.

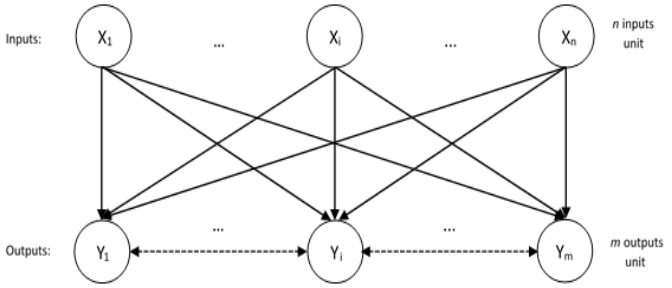


Fig.2. SOM Architecture

In this study, the SOM algorithm steps as follow:

1. Generate training data P_{nxk}
2. Set parameters of the cluster (m), initialize weights W_{mxk}, initialize member cluster G_{mxk} and the learning rate (η)
3. ***For*** input vector k
   Each i, calculate ED (L^2)
   Select Best Matching Unit (min)
   Update weights, learning rate (η)
   End
4. Repeat steps 2 until no point changes its cluster assignment, or until a maximum number of passes through the data set is performed.

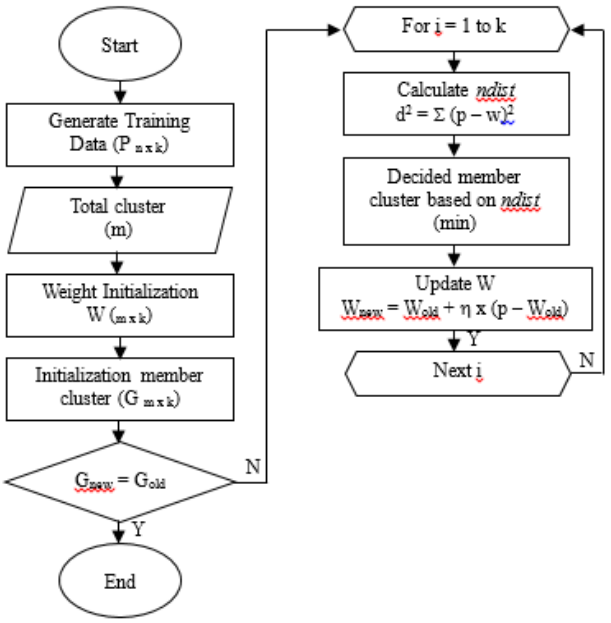


Fig.3.SOM Flowchart

2.4. DATASETS

Mahakam River covers an area of 86.024 km2, including Samarinda, Balikpapan, Kutai Barat, Kutai Kartanegara, Malinau and Penajam Paser Utara. In this study, we use yearly rainfalls that are obtained from the 13 stations. The yearly rainfalls are recorded in 1986-2008 and taken as a sample dataset. The min, max, and average are used as training data patterns. Before the data analysis process, the dataset should be normalized. The normalized dataset is shown in Table 1 and Fig. 4. Then, the dataset are analyzed by using a MATLAB R2013b.

Table.1. Real Yearly Rainfall Dataset 1986-2008 in minimum, maximum, and average values

NO	OBSERVATION AREA	MIN	MAX	AVG
1	L. Iram	1057	4598	2814
2	Melak	0	4273	1724
3	K. Bangun	1346	4037	2366
4	Ma. Kaman	0	7692	3923
5	T. Dalam	955	3051	2054
6	Tenggarong	797	7263	3586
7	Ma. Ancalong	396	1611	1072
8	Temindung	1566	2758	2218
9	Baqa	0	3828	1967
10	Samboja	112	2859	1494
11	Klandasan	744	3654	2082
12	Sepinggan	1483	3786	2690
13	Waru	637	3296	2126

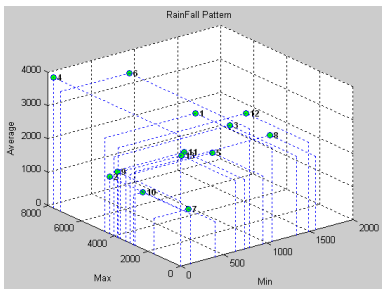


Fig.4. Plot of training data in min, max, and average

2.5. PERFORMANCE OF MEASUREMENT

The most important component of the cluster algorithm is the measure distance computed between the data and the centroid. Several “distance-space” methods have been implemented in calculating the distance between them. In this study, the Euclidean distance is used to calculate the distance. In principle, the Euclidean distance is to calculate the distance in the “distance-space” by calculating two points in the shortest distance.

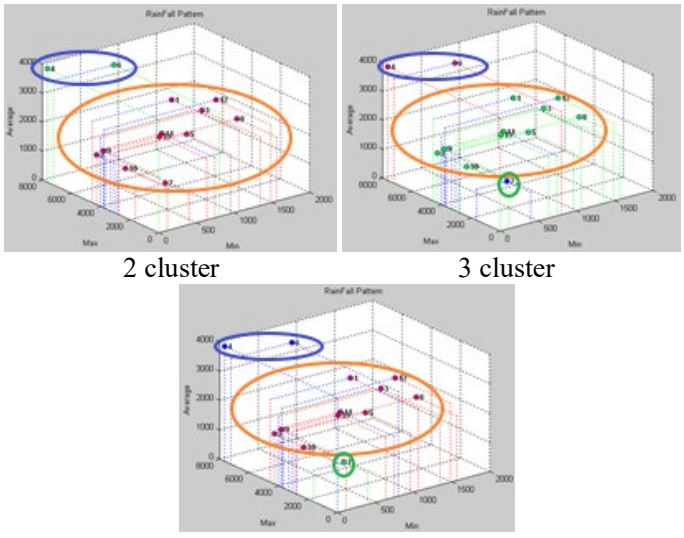
4. EXPERIMENTAL RESULT

This section presents the results obtained for the *k*-Means and SOM algorithms in clustering yearly rainfall datasets. In order to evaluate the performance of clustering dataset, *ndist* method is implemented.

4.1. RESULT OF K-MEANS ALGORITHM

In this experiment, training datasets have been classified into 2, 3, and 4 clusters. According to the basic

k-mean clustering algorithm, clusters are fully dependent on the selection of the initial clusters centroids. *k* data elements are selected as the initial centers; then distances of all data elements are calculated by using the Euclidean distance formula. In this experiment, an average data is used as one of centroid value. Then, another centroid is determined randomly in order to observe good patterns of groups. The results of *k*-Means are shown in Figure 5.



2 cluster 3 cluster

4 cluster

Fig.5. *k*-Means Algorithm Results

Based on experimental results obtained, clusters with *k*=3 and *k*=4 produce the same number of clusters, which is 3 clusters. In this experiment, the grouping of the 3 clusters is selected because if the centroid value were randomly selected (e.g., for 3 cluster), there will be inconsistent results. This proves that one should have the centroid as the average of all the training data. The results are shown in the Fig. 6.

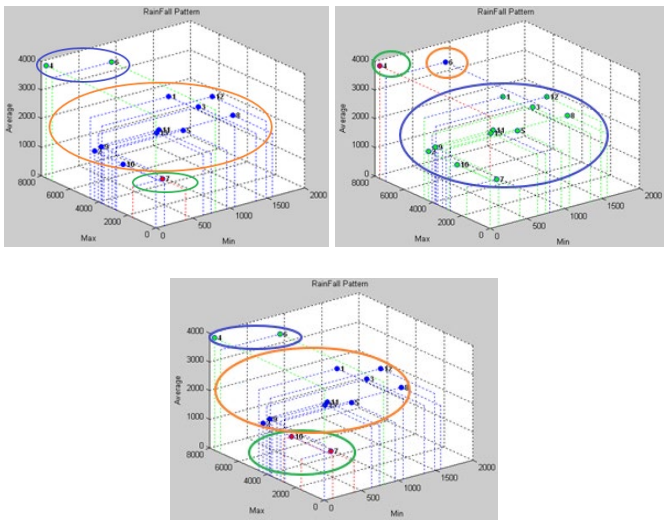
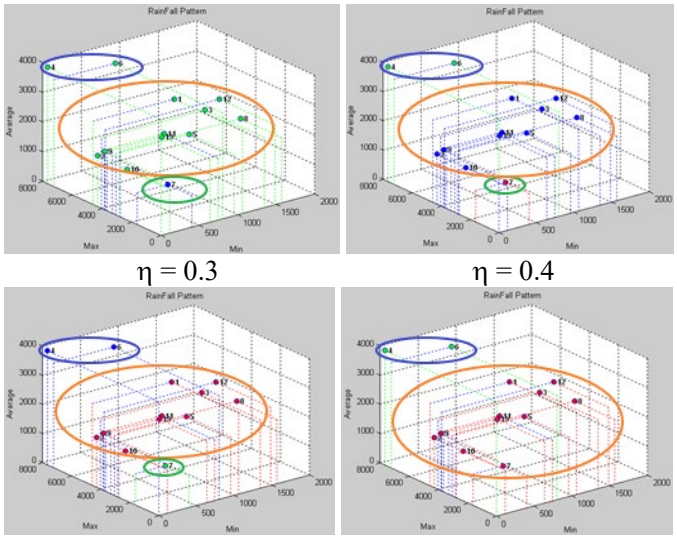


Fig.6.K-Means Algorithm Inconsistent Results

4.2. RESULT OF SOM ALGORITHM

In the SOM technique, the initialization weighted intra-layer has been randomly settled. The weighting matrix of size *m* is the number of clusters and *l* is the number of attribute data patterns training. In this experiment, three clusters were formed using the 2 x 3 dimension. Then, centroid was calculated by using the Euclidean Distance and grouping all data in order to compare the clustering process using the *k*-Means algorithm. Meanwhile, learning rate (η) 0.3, 0.4, 0.5, 0.8, and 0.9 were tested. The results are shown in the following Fig.7.



η = 0.3 η = 0.4

η = 0.8 η = 0.9

Fig.7. SOM Clustering Results

Based on the experimental, it appears that for learning rate η = 0.4; η = 0.8 produces the same cluster members. These results were also the same when using the *k*-Means algorithm as shown in Fig. 8 and Table 2.

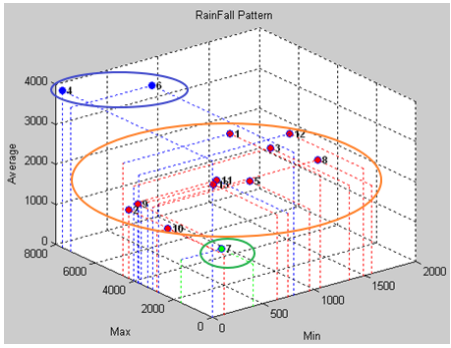


Fig.8. K-Means and SOM Clustering Results

Table.2. Clustering Observation Areas Results

Clusters	Observation Areas
1	Long Iram, Melak, Kota Bangun, Teluk Dalam, Temindung, Baqa, Samboja, Klandasan, Sepinggan, Waru
2	Muara Ancalong
3	Muara Kaman, Tenggarong

5. CONCLUSIONS

In this study, the *k*-Means and SOM with a hierarchical cluster analysis were systematically applied for clustering rainfall datasets from 13 observation areas. Based on the results obtained, the Euclidean distance method is capable of grouping similar data into the same unit. *k*-Means clustering algorithm is found to be able to work on the assumption that the initial centers are provided. The search for the final clusters or centers starts from these initial centers. The *k*-Means algorithm is a simple and fast algorithm that performs well in analyzing hydrology data. In order to get the optimal distance, the parameters that need to be considered in the algorithm *k*-Means includes the number of clusters *k*, cluster initialization, and distance metric. However, the weakness of the *k*-Means clustering algorithm is that it often produces different results (local optima). In contrast, SOM has several characteristics, i.e., SOM could visualize the grouping result in two-dimensional topography form, so the observation of the distribution group will be easier. The SOM algorithm also requires a description function, learning rate, the number of groups, and the number of iterations. Thus, a trial-and-error with some parameter values is required before the best parameter values are selected. SOM is relatively suited for data that has the number of the group known by observing the natural shape of the data distribution. Further works need to focus on identifying the optimal clustering models for rainfall conditions in Mahakam River based on heuristically approaches.

REFERENCES

[1]. Mohd Nasir Mat Amin, Puteri N. E. Nohuddin, and Zuraini Zainol, 'Trend Cluster Analysis Using Self Organizing Maps', in *2014 Fourth World Congress on Information and Communication Technologies (WICT)©* 2014 IEEE, 2014), pp. 80 - 84.

[2]. Aymen Cherif, Hubert Cardot, and Romuald Boné, 'Som Time Series Clustering and Prediction with Recurrent Neural Networks', *Neurocomputing,* 74 (2011), 1936–44.

[3]. Aditya Dubey, and Sanjiv Sharma, 'Performance Measurement of K-Means & Spectral Clustering Graph Clustering Algorithms', *International Journal of Advanced Computer Science,* Vol. 4, No. 12, Dec., 2014. (2014), 565-70.

[4]. Haviluddin, Rayner Alfred, Joe Henry Obit, Mohd Hanafi Ahmad Hijazi, and Ag Asri Ag Ibrahim, 'A Performance Comparison of Statistical and Machine Learning Techniques in Learning Time Series Data', *Advanced Science Letters* (2015), 3037-41.

[5]. Anil K. Jain, 'Data Clustering: 50 Years Beyond K-Means', in *19th International Conference on Pattern Recognition (ICPR)* (Tampa, FL, 2008), pp. 1-33.

[6]. A.M. Kalteh, P. Hjorth, and R. Berndtsson, 'Review of the Self-Organizing Map (Som) Approach in Water Resources: Analysis, Modelling and Application', *Environmental Modelling & Software,* 23 (2008) (2008), 835-45.

[7]. Teuvo Kohonen, 'Matlab Implementations and Applications of the Self-Organizing Map', 2014).

[8]. Mislan, Haviluddin, Sigit Hardwinarto, Sumaryono, and Marlon Aipassa, 'Rainfall Monthly Prediction Based on Artificial Neural Network: A Case Study in Tenggarong Station, East Kalimantan - Indonesia', in *International Conference on Computer Science and Computational Intelligence (ICCSCI 2015)* (Binus, Jakarta: © 2015 The Authors. Published by Elsevier B.V., 2015), pp. 142 – 51.

[9]. Thuy Thanh Nguyen, Akira Kawamura, Thanh Ngoc Tong, Naoko Nakagawa, Hideo Amaguchi, and Gilbuena Jr. Romeo, 'Clustering Spatio–Seasonal Hydrogeochemical Data Using Self-Organizing Maps for Groundwater Quality Assessment in the Red River Delta, Vietnam', *Journal of Hydrology,* 522 (2015) (2015), 661–73.

[10]. Sharifah Mohd. Sharif, Faradiella Mohd. Kusin, Zulfa Hanan Asha'ari, and Ahmad Zaharin 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 The Authors. Published by Elsevier B.V., 2015), pp. 73 – 78.

[11]. Archana Singh, Avantika Yadav, and Ajay Rana, 'K-Means with Three Different Distance Metrics', *International Journal of Computer Applications,* 67– No.10, April 2013 (2013), 13-17.

[12]. Y. S. Thakare, and S. B. Bagal, 'Performance Evaluation of K-Means Clustering Algorithm with Various Distance Metrics', *International Journal of Computer Applications,* 110 – No. 11, January 2015 (2015), 12-15.

[13]. Yun Xie, Shui-qing Yin, Bao-yuan Liu, Mark A. Nearing, and Ying Zhao, 'Models for Estimating Daily Rainfall Erosivity in China', *Journal of Hydrology,* 535 (2016) (2016), 547–58.

[14]. Zahra Zahmatkesh, Mohammad Karamouz, and Sara Nazif, 'Uncertainty Based Modeling of Rainfall-Runoff: Combined Differential Evolution Adaptive Metropolis (Dream) and K-Means Clustering', *Advances in Water Resources,* 83 (2015) (2015), 405–20.

[15]. National Centers for Environmental Information, Accessed 27 June 2016, http://www.ncdc.noaa.gov/cdoweb/search

[16]. National Water Information System: Web Interface (USGS 01302020 Bronx River at NY Botanical Garden at Bronx NY), Accessed 27 June 2016, http://waterdata.usgs.gov/usa/nwis/uv?01302020

Received: 22 September 2010. Accepted: 18 October 2016