K-Affinity Propagation (*K-AP*) and K-Means Clustering for Classification of Earthquakes in Indonesia

Muhammad Muhajir, Nilam Novita Sari
Department of Statistics
Islamic University of Indonesia
Yogyakarta, Indonesia
mmuhajir@uii.ac.id, nilamnovitasari2013@gmail.com

***Abstract*—Indonesia is located among three plate tectonics and ring of fire cause to be disaster prone especially to the earthquakes. Indonesia is the state of being having a level by earthquake high in the world as many as fifty nine thousand and eighty nine the earthquakes was happening in 2017.Hence, its efforts to mitigate is needed to reduce the impact caused by the earthquakes. Clustering K-Affinity Propagation and K-Means method used to cluster areas the earthquake, clustering aims to its efforts to mitigate. Testing validation use C-Index, Davies Bouldin Index, and Connectivity Index whereby the test was constructed using three and five clusters for K-Means and was constructed using two and four clusters using K-AP. Based on clusters of variance got that clusters of 4 is the sum clusters best has value ratio of the smallest. The results obtained the number of clusters one has two members with exemplar Celebes Sea, cluster two has twelve members with exemplar Halmahera, cluster three has onemember with exemplar Minahassa Peninsulaand cluster four has thirty-four members with exemplar Sumba Region.**

Keywords—earthquakes; K-Affinity Propagation (K-AP);clustering;K-Means

I. INTRODUCTION

Indonesia is the largest archipelago which have more than thirteen thousand islands that lie between the three plate tectonics deeds Eurasian plate, indo-Australia plate, and plate of the Pacific Ocean as well as in the ring of fire that causes Indonesia to natural disaster prone. Natural disasters could not be avoided because many factors are cause to happen, one of them is a factor of geology. One of the disasters caused by geological factors is earthquake.

Earthquake is a disaster which is formed by the vibrations on the earth surface. According to BNPB in 2012, the data shows that Indonesia was a country has the high level of seismicity in the world. Data from BMKG shows that 5989 earthquakes happened in Indonesia in 2017, that means about 17 earthquakes happen every day as well as the strength of earthquake is about 3.3 SR till 7.2 SR in 2017.

Most areas of Indonesiawere the area with the most earthquake prone, only a few regions in Indonesia unaffected by the threat of an earthquake. As a result of the people living in the areas prone to the earthquake have to always stay alert because an earthquake could come anytime. Because of that, its efforts to mitigate is needed in order to reduce the impact of due to a disaster including readiness in at risk of long term.

In conducting its efforts to mitigate earthquake, can be done clustering against Indonesian regions affected areas earthquake, in order to assist the government for its efforts to mitigate areas prone to earthquakes.In this paper, K-Affinity Propagation(K-AP) and K-Meansare comparedto decide the best method according to data accuracy and usefulness.

K-AP method is a methodof Affinity Propagation (AP)are modified to produce the optimal number of exemplar through AP.K-AP method is a new cluster method which identify exemplar between all data points, and form cluster from the data points around the exemplar [1].K-Means is a way of non-hierarchical grouping which tries to separate data into 1 or more categories. K-Means algorithm starts with a random selection of K, then set the values of K randomly and these values will be the center of the grouping or cluster. The next step calculates the distance of every data to each center point, then categorize each data based on this calculation [2].

From the background, we propose application of K-Affinity Propagation and K-Means method on the occurrence of earthquake in Indonesia. The both methods analysis has to be done to make restitution of the clusters of static where there was no change in the exemplar, so as to produce the number of cluster unchanged every testing.

II. RELATED WORK

This paper used some previous paper as a reference, first is paper from Zhang et al. entitled K-AP: Generating Specified K Clusters by Efficient Affinity Propagation, this paper compared the distortion and computational cost between K-AP, AP and K-Medoids. The result of this paper were obtained that K-AP more efficient than APin terms of computational cost in generating specified K clusters while more effective than K-Medoids in terms of distortion minimization [3].

Second paper is from Serdah and Ashour entitled Clustering Large-Scale Data Based on Modified Affinity

Propagation Algorithm where the purpose of this paper is to produce the best cluster for large scale data. This paper used 2 methods that is K-AP and IWC (Interdepartmental Water Clustering). Data points are clustered into small groups, then applying K-AP method and the applied IWC to find the worldwide models for original matching clusters. The result of this paper were obtained that K-AP – IWC is the best clustering method for large scale data than AP, K-AP and Hierarchical Affinity Propagation [4].

And third paper is a paper entitled Analysis and Implementation of Algorithm Clustering Affinity and K-Means at Data Students Based on GPA and Duration of Bachelor-Thesis Completion. This paper is compare K-Means and Affinity method to define the best cluster algorithm in terms of accuracy and utility. The result of this paper were obtained that Affinity method was the best cluster method because its data cluster results are more accurate and effective than K-Means. Affinity method gives stable cluster results where the value of affinity propagation exemplar has remained constant even after five trials. In contract, K-Means gives different values of its centroid after every trials [2].

The next is a paper entitled Implementation of K-Means Clustering Method for Electronic Learning Model. The purpose of this paper is to group students' learning activities using e-learning where the clustering was conceived by simulating 396 students' activities, namely student classroom participation, submission of assignments, viewing of assignments, increase in discussions and comment, downloading of course materials, viewing of articles and tests, and test submissions where activities of 10 sample students were observed. The result of this paper obtained 2 clusters which are cluster of students' activity and improvement of student's ability, where cluster 1 has membership percentage of 70% and cluster 2 has membership percentage of 30% [5].

The fifth paper from Singh and Kaur entitled Comparison Analysis of K-Means and Kohonen-SOM Data Mining Algorithms Based on Student Behaviors in Sharing Information on Facebook. This paper is compared K-Means and Kohonen-SOM using datasets "information sharing in Facebook" with approximately 10 descriptors and 1850 instances. The result of this paper obtained that Kohonen-SOM gives the better performance than K-Means with minimum rate of error (high accuracy), minimum computation time based on the same data set and parameters [6].

III. THEORICAL BASIS

A. Affinity Propagation (AP)

AP is an algorithm that identifies exemplars among data points and forms clusters of data points around these exemplars. It works by simultaneously considering all data points as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters arises. Exemplar is a data point that is a good representative of itself and other data points [1].

Affinity propagation assumes an input of a collection of real-valued similarities between data points, where the similarities (*i, k*) indicates how well the data point with index k is suited to be the exemplar for data point *i*[1]. The algorithm of Affinity Propagation was calculated using (1)–(7) [4]:

- Input Similarities

$$s(i,k) = \sum -||x_i - x_k||^2 \quad (1)$$

where

s : similarity matrix

i , k : 1, 2, …, n

- Initialize availability matrix to zero

$$a(i,k) = 0,\ i \in \{1, 2, \ldots, n\} \quad (2)$$

- Update responsibilities:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\} \quad (3)$$

Update self-responsibilities:

$$r(j,j) = s(k,k) - \max\{s(i,k')\} \quad (4)$$

- Update matrix availabilities:

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\}\} \quad (5)$$

Update self-availability:

$$a(k,k) = \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\} \quad (6)$$

Until converge

- Combined availability and responsibility:

$$c(i,k) = a(i,k) + r(i,k) \quad (7)$$

B. K-Affinity Propagation

K-AP was modified to create a given number of an optimal set of exemplars through Affinity Propagation [4]. By adding a rule or control in the process of message passing to restrict the number of clusters to be K, while keeping all AP advantages in clustering, K-AP can generate K clusters based on the user's needs and parameters. Another advantage of K-AP over AP is the confidence in one data item to be an exemplar which is automatically self-adapted by K-AP while the confidence in AP is a parameter set by a user. Moreover, the overhead computational cost for K-AP is insignificant as compared to AP. However, similar to AP, the limitations of clustering large-scale data still exists. It still consumes time and memory while processing large-scale data [4]. The algorithm of K-Affinity Propagation was calculated using (8)–(17) [3]:

- Input similarities

$$\{s(i,k)\}_{i,k \in \{1,\ldots,N\} i \neq k}, K \quad (8)$$

- Initialize availabilities and confidence matrix to zero.

$$a(i,k) = 0 \quad (9)$$

$$\eta^{out}(i) = \min(s) \quad (10)$$

- Update responsibilities:

$$r(i,k) = s(i,k) - \max\{\eta^{out}(i) + a(i,i), \max_{k':k' \in \{i,k\}}\{a(i,k') + s(i,k')\} \quad (11)$$

Update self-responsibility:

$$r(i,i) = \eta^{out}(i) - \max_{k':k' \neq i}\{a(i,k') + s(i,k')\} \quad (12)$$

- update availabilities matrix:

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\}\} \quad (13)$$

Update self-availability:

$$a(k,k) = \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\} \quad (14)$$

- Update confidence:

$$\eta^{in}(i) = a(i,i) - \max_{k':k' \neq i}\{a(i,k') + s(i,k')\} \quad (15)$$

$$\eta^{out}(i) = -R^{K}(\{\eta^{in}(j), j \neq i\}) \quad (16)$$

Until converge

- Combined availability and responsibility:

$$c(i,k) = \text{argmax}_j\{a(i,k) + r(i,k)\} \quad (17)$$

C. K-Means

K-Means is a method of non-hierarchical data cluster which tries to categorize data into one or more cluster based on similar characteristics. K-Means is also referred to as the repeatedly clustering algorithm. K-Means algorithm starts with a random selection of k, where k is the number of clusters to be formed. Then set the values of k randomly, for a while these values will be a center of cluster or commonly called centroid, mean, or means. To find the closest distance of each data with the centroid, calculate the proximity of every data to each centroid using Euclidean formula. Then, classify each data based on data to the centroid. The data point will always migrate until the centroid does not change again (stable) [2].

K-Means is a method of examining data or methods that perform data mining modeling process without human supervision and it is one method that classifies data with the partition system. K-Means method is trying to categorize data into several groups, where each group have different characteristics. In other words, this algorithm seeks to minimize the variation among the data in one cluster and maximize the variation with data on other cluster [2].

D. Validation Test

- The C-Index calculated using (18)[7]:

$$C - \text{Index} = \frac{S_w - S_{min}}{S_{max} - S_w}, S_{min} \neq S_{max} \in (0,1) \quad (18)$$

where:

$$S_W = \sum_{k=1}^{q} \sum_{\substack{i,j \in C_k \\ i<j}} d(x_i, x_j) \quad (19)$$

The minimum value of the index is used to indicate the optimal number of clusters.

- Davies and Bouldinindex calculated using (20)–(21)[7]:

$$DB_{(q)} = \frac{1}{q}\sum_{k=1}^{1} max_{k \neq 1}\left(\frac{\delta_k + \delta_l}{d_{kl}}\right) \quad (20)$$

where:

$$\delta_k = \sqrt[u]{\frac{1}{n_k}\sum_{i \in C_k}\sum_{j=1}^{p} |X_{ij} - C_{kj}|^u} \quad (21)$$

δ_k:size of cluster disperse of C_k(for u=2, δ_k is st. dev from object distance on the C_k cluster to cluster centers.

k, l : 1, …, q = the number of cluster

d_{kl}:Euclidean distance

The minimum value of the index is used to indicate theoptimal number of clusters.

- Connectivity Index calculated using (22)[8]:

$$Conn(C) = \sum_{i=1}^{N}\sum_{k=1}^{L} X_{i,nn_{i(j)}} \quad (22)$$

where:

$nn_{i(j)}$: nearest neighbor observation *i* to *j*

L: parameter that determine the number of neighbors who contribute to the measurement of connectivity

The minimum value of the index is used to indicate theoptimal number of clusters.

E. Determine Goodness of The Cluster Method

After the classification process, clustering results of the three methods was further assessed. The assessment is done by comparing the sum of squares within-group and the sum of squares between-group. The best method is one that has the smallest ratio of standard deviation [9]. The standard deviation in the group (S_w) can be calculated using (23).

$$S_w = K^{-1}\sum_{k=1}^{K} S_k \quad (23)$$

where:

K: number of cluster formed

S_k:standard deviation of k *cluster*

The standard deviation betweencluster (S_b) can be calculated using (24).

$$S_b = [(K-1)^{-1}\sum_{k=1}^{K}(\bar{X}_k - \bar{X})^2]^{1/2} \quad (24)$$

where:

S_b: standard deviation between cluster

$\bar{X}_k$: mean of k cluster

$\bar{X}$: the overall mean of cluster

The best method is a method who has the lower ratio of S_w/S_b.

IV. RESEARCH METHOD

This paper used earthquake data in Indonesia in 2017. To address the current problems, can be done by compare between K-Meansand K-AP using 3 index cluster validity, that is C-Index, Davies Bouldin Index and connectivity Index. Based on the workflow stages,troubleshooting on K-Means and K-AP consisted of several stages of problem as seen in Fig. 1.

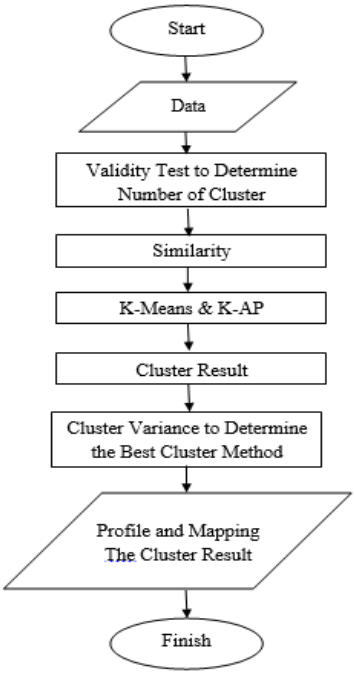

Fig. 1. Flowchart cluster method

V. RESULTS

Table I shows validation test for K-Means and K-AP using C-Index, Davies Bouldin Index and Connectivity Index. Validation test was obtained the optimal number of cluster for K-Means was 3 and 5, meanwhile the optimal number of cluster for K-AP was 2 and 4. So, to decide the optimal number of cluster, it'll be used 3 and 5 cluster for K-Means, 2 and 4 cluster for K-AP for clustering earthquakes in Indonesia, and then it will compare the best results of the clusters.

TABLE I. VALIDATION TEST

Cluster Methods	K	C-Index	DaviesBouldin	Connectivity
K-MEANS	2	0.102	0.658	8.247
	3	0.060	0.109	5.102
	4	0.067	0.113	9.419
	5	0.052	1.651	24.531
	6	0.061	1.212	30.049
K-AP	2	0.118	0.642	2.173
	3	0.060	0.109	5.102
	4	0.022	0.108	8.960
	5	0.052	1.651	24.531
	6	0.050	1.483	28.283

Table II shows the cluster results using K-Means with the number of clusters 3, where for 1st cluster has 12 members, 2nd cluster has 1 member and 3rd cluster has 25 members. The members of each clusters can be seen in the Table II.

TABLE II. CLUSTER RESULT OF K-MEANS USING 3 CLUSTERS

Cluster	Members
1	Bali Region, Banda Sea, Halmahera,Irian Jaya Region, Northern Molucca Sea, South of Java, Sumbawa Region, Java, Northern Sumatra,Seram, Southern Sumatra, Sulawesi
2	Minahassa Peninsula
3	Arafura Sea, Aru Islands Region,BaliSea,Borneo, Buru, Celebes Sea, Ceram Sea, Flores Region,FloresSea,Irian Jaya, Java Sea, Near North Coast of Irian Jaya,Northof Halmahera, Off West Coast of Northern Sumatra, Savu Sea, South of Bali, South of Sumbawa, Southern Molucca Sea, Southwest of Sumatra, Sumba Region, Sunda Straits, Talaud Islands, Tanimbar Island, Timor Region, Timor Sea,

Table III shows the cluster results using K-Means with number of cluster 5 where for 1st cluster has 9 members, 2ndcluster only has 1 member that is Minahassa Peninsula, 3rd cluster has 18 members, 4th cluster has 8 members and 5th cluster has 2 members. The members of each clusters can be seen in the Table III.

TABLE III. CLUSTER RESULT OF K-MEANS USING 5 CLUSTERS

Cluster	Members
1	Ceram Sea, Flores Region,Irian Jaya Region, Near North Coast of Irian Jaya,Northern Molucca Sea, South of Java, Sumbawa Region, Talaud Islands, Timor Region
2	Minahassa Peninsula
3	Arafura Sea, Aru Islands Region,BaliSea,Borneo, Buru, Flores Sea,Irian Jaya, North of Halmahera, Off West Coast of Northern Sumatra, Savu Sea, South of Bali, South of Sumbawa, Southern Molucca Sea, Southwest of Sumatra, Sumba Region, Sunda Straits, Tanimbar Island, Timor Sea
4	Bali Region, Banda Sea, Halmahera, Java, Northern Sumatra, Seram, Southern Sumatra, Sulawesi
5	Celebes Sea, Java Sea

Table IV shows the cluster results using K-AP where 1st cluster has 30 members with exemplar Irian Jaya Region and 2nd cluster has 9 members with exemplar Java. The members of each cluster can be seen in the Table IV.

TABLE IV. CLUSTER RESULT OF K-AP USING 2 CLUSTERS

Cluster	Exemplar	Members
1	Irian Jaya Region	Arafura Sea, Aru Islands Region,BaliSea,Borneo, Buru, Celebes Sea, Ceram Sea, Flores Region,FloresSea,Irian Jaya Region, Irian Jaya, Java Sea, Near North Coast of Irian Jaya, North of Halmahera, Northern Molucca Sea, Off West Coast of Northern Sumatra, Savu Sea, South of Bali, South of Java, South of Sumbawa, Southern Molucca Sea, Southwest of Sumatra, Sumba Region, Sumbawa Region, Sunda Straits, Talaud Islands, Tanimbar Island, Timor Region, Timor Sea
2	Java	Bali Region, Banda Sea, Halmahera, Java, Minahassa Peninsula, Northern Sumatra, Seram, Southern Sumatra, Sulawesi

Table V shows the cluster results using K-AP for where 1st cluster has 2 members with exemplar Celebes Sea, 2nd cluster has 12 members with exemplar Halmahera, 3rd cluster has 1 member that is Minahassa Peninsula whereby the Minahassa Peninsula be the exemplar for 3rd cluster and 4th cluster has 24

members with exemplar Sumba Region. The members of each clusters can be seen in the Table V.

TABLE V. CLUSTER RESULT OF K-AP USING 4 CLUSTERS

Cluster	Exemplar	Members
1	Celebes Sea	Celebes Sea, Java Sea
2	Halmahera	Bali Region, Banda Sea, Halmahera, Irian Jaya Region,Northern Molucca Sea, South of Java, Sumbawa Region, Bali Region, Banda Sea, Halmahera,Java, Northern Sumatra, Seram, Southern Sumatra, Sulawesi
3	Minahassa Peninsula	Minahassa Peninsula
4	Sumba Region	Arafura Sea, Aru Islands Region,BaliSea,Borneo, Buru, Flores Sea,Irian Jaya, Near North Coast of Irian Jaya,North of Halmahera, Ceram Sea, Off West Coast of Northern Sumatra, Savu Sea, Flores Region, South of Bali, Southern Molucca Sea, Southwest of Sumatra, Sumba Region, Sunda Straits, Talaud Islands, Tanimbar Island, Timor Region, Timor Sea South of Sumbawa

Table VI shows goodness of the cluster method where the low value of Sw shows the result better, while for Sb, the high value shows the result better. To decide the best number of cluster, can be used the lower ratio of Sw/Sb. From the Table VI The best cluster is using K-AP with 4 clusters cause' it has lower ratio of Sw/Sb.

TABLE VI. GOODNESS OF THE CLUSTERS METHOD

Methods	Cluster	Sw	Sb	Sw/Sb
K-Means	3	10.379	73.538	0.141
	5	6.701	57.139	0.117
K-AP	2	25.204	37.303	0.676
	4	6.652	62.252	0.107

Fig. 2 was maps result from K-AP clustering where the members of 1st cluster was given coordinates in Black, the members of 2nd cluster was given coordinates in yellow, the member of 3rd cluster was given coordinate in green and the members of 4th cluster was given coordinates in blue.

Fig. 2. Maps result using best cluster method.

VI. CONCLUSION

From the result were obtained some conclusion, that is validation test result using C-index, Davies Bouldin Index and Connectivity Index for K-AP was obtained the optimal number of cluster 2 and 4, meanwhile for K-Means was obtained the optimal number of cluster 3 and 5. Using cluster variance it got K-AP using 4 cluster is the best method because it has the lower value of Sw/Sb. The region in 3rd cluster should be supervised because in the 3rd cluster is a the most frequent area of the earthquakes.

REFERENCES

[1] B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," *Science (80-.).*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.

[2] R. Refianti, A. B. Mutiara, A. Juarna, and S.N. Ikhsan, "Analysis and Implementation of Algorithm Clustering Affinity and K-Means at Data Students Based on GPA and Duration of Bachelor-Thesis Completion," *Journal of Theorical an Applied Information Technology*, vol. 35 no.1,2015.

[3] X. Zhang, W. Wang, K. Norvag, and M. Sebag, "K-AP: Generating Specified K Clusters by Efficient Affinity Propagation," in *2010 IEEE International Conference on Data Mining*, 2010, pp. 1187–1192.

[4] A. M. Serdah and W. M. Ashour, "Clustering large-scale data based on modified affinity propagation algorithm," *J. Artif. Intell. Soft Comput. Res.*, vol. 6, no. 1, pp. 23–33, 2016.

[5] H. L. Sari, D. Suranti, and L. N. Zulita, "Implementation Of K-Means Clustering Method For Electronic Learning Model," *J. Phys. Conf. Ser.*, vol. 930, no. 1, p. 012021, Dec. 2017..

[6] G. Singh, and A.Kaur, "Comparative Analysis of *K-Means* and Kohonen-SOM Data Mining Algorithms Based on Student Behaviors in Shaaring Information on Facebook," *International Journal of Engineering and Computer Science*, vol. 6, issue 4, 2017.

[7] M. Charrad,N. Ghazallii,V. Boiteau, and A.Niknafs, "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set",*Journal of Statistical Software*, vol. 61, issue 6, 2014

[8] G.Brock,V.Pihur, S.Datta, and S.Datta, "clValid: An R Package for Cluster Validation,"*Journal of Statistical Software*,vol. 25, issue 4, 2008.

[9] M. J. Bunkers, J. R. Miller, A. T. DeGaetano, M. J. Bunkers, J. R. M. Jr., and A. T. DeGaetano, "Definition of Climate Regions in the Northern Plains Using an Objective Cluster Modification Technique," *J. Clim.*, vol. 9, no. 1, pp. 130–146, Jan. 1996.