

An improved clustering based on K-means for hotspots data

Rani Rotul Muhima¹, Muchamad Kurniawan¹, Septiyawan Rosetya Wardhana¹, Anton Yudhana², Sunardi², Mitra Adhimukti³

¹Department of Informatics Engineering, Faculty of Electrical Engineering and Information Technology, Institut Teknologi Adhi Tama Surabaya (ITATS), Surabaya, Indonesia

²Department of Electrical Engineering, Faculty of Industrial Technology, Ahmad Dahlan University, Yogyakarta, Indonesia

³Prevention Division, Regional Disaster Management Agency of Riau Province, Pekanbaru, Indonesia

Article Info

Article history:

Received Nov 25, 2022

Revised Apr 12, 2023

Accepted Apr 16, 2023

Keywords:

Cluster center

Clustering

Genetic algorithm polygamy

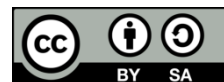
Hotspot data

K-means

ABSTRACT

Riau province is one of the provinces in Indonesia where forest fires frequently occur every year. Hotspot data is geothermal points and they can be utilized as an indicator of forest fires. Clustering's method can be used to analyze potential forest fires from hotspot data's cluster pattern. In this study, hybrid genetic algorithm polygamy with K-means (GAP K-means) was used for hotspot data clustering. GA polygamy was used to determine the initial centroid of K-means. It was used to solve the sensitivity of K-means to the initial centroid, and to find the optimal solution faster. Experimentally compared the performance of GAP K-means, GA K-means, and K-means on the hotspots data, two artificial datasets, and three real-life datasets. Sum square error (SSE), davies bouldin index (DBI), silhouette coefficient (SC) and F-measure are used to evaluation clustering. Based this experiment, GAP K-means outperforms than K-means but GAP K-means still not fast to achieve convergent than GA K-means.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Muchamad Kurniawan

Department of Informatics Engineering, Faculty of Electrical Engineering and Information Technology

Institut Teknologi Adhitama Surabaya (ITATS)

Surabaya, Indonesia

Email: muchamad.kurniawan@itats.ac.id

1. INTRODUCTION

Riau is one of the provinces of Indonesia where forest fires frequently occur every year. From 1998 until 2016, cases of forest and peat fires in Riau were status emergency smoke [1] and in 2021 it was recorded that 14,939 hectares were burned [2]. The effects of forest fires, such as haze were felt not only in Riau, but also in other provinces and even neighboring country such as Malaysia and Singapore. Forest fires' losses also impacted in multiple sectors, including economy [1], [3], public health, livelihood, and land degradation [4]. It is essential to detect and predict the occurrence of forest fires for prevention. This problem can be solved by analyzing the distribution of hotspot data. Hotspot data is geothermal points and they can be utilized as an indicator of forest fires [5]. Clustering is an important analytical technique. It can be used to analyze potential forest fires from hotspot data's cluster pattern.

Clustering is one of data mining technique. Clustering divides a data set into several groups based on the similarity of objects. The similarity is maximized with objects in the same cluster and with objects from different clusters, similarity is minimized [6]. Clustering algorithms can be classified two categories: hierarchy clustering and partitional clustering. In hierarchy clustering we find density based algorithm [7], graph-based algorithm [8], hybrid algorithm [9] and prototype-based algorithm [10]. In partitional clustering, we find well-

known traditional method, such as K-means, simulated annealing [11] and fuzzy c-means [12]. The most popular of partitional clustering is K-means.

K-means is commonly used today [13] because of simplicity [14], high clustering speed [15] and easy to implement. However, K-means has drawback that sensitive to initial centroid (cluster center) [14], [16]-[20]. The quality of the initial centroids influences the clustering quality [14], [18]. Several studies were conducted in order to improve K-means' initial centroid quality. Among of them is done using the optimization method, such as genetic algorithm (GA) [14], particle swarm optimization (PSO) [21], firefly algorithm (FA) [19], hybrid GA-PSO-and fuzzy system [20]. Their proposed methods are superior to conventional K-means clustering.

The genetic algorithm as a new global optimization search algorithm is simple to implement, robust and become an important intelligent algorithms [22]. Maulik dan Bandyopadhyay [23], for optimizing the similarity metrics of cluster, the GA method is used to find center of cluster. Their proposed method shows better performance than K-means on artificial and real-life dataset. Rahman and Islam [14] proposed clustering that combination of GA with K-means and called GenClust. GA is used to determine the initial centroid of the K-means process. In GenClust, initial genes on the GA process were determined. Their experiment results show their method is superior to some existing techniques.

Obtaining GA convergence sometimes takes a long time [24], [25]. Aibinu *et al.* [24] proposed GA polygamy clustering for route optimization. And the results showed that their algorithm better than some existing techniques and time to reach convergence is faster. GA polygamy clustering outperforms GA clustering based on sum square error (SSE) values and the convergence time is faster than GA clustering [26]. This advantage of GAP is used to improve GA process to determine initial centroid in GA K-means and called genetic algorithm polygamy with K-means (GAP K-means).

This paper proposed hybrid GAP K-means for clustering hotspot data. GA polygamy is used to determined initial centroid of K-means, which is used to solve the sensitivity of K-means to the initial centroid. And in order to find optimal solution faster. Experimental result will be compared the performance of GAP K-means, GA K-means and K-means on the hotspots data, two artificial dataset and three real-life datasets. We also compare four evaluation clustering: SSE, Davies Bouldin index (DBI), silhouette coefficient (SC), and F-measure.

Main contribution of this paper is development of the GA K-means method which was previously able to overcome problem of K-means. The problem is related to the determination of initial centroids. The development devoted to determination of the initial centroid. The mating of one father with more than one mother is expected to speed up process of finding the optimal initial centroid. So, GAP K-means is expected to have better performance with time to converge faster than GA K-means. The rest of paper is organized as follow: section 2 describes data and methods used in this study. Section 3 presents result and discussion, where the performance of the proposed clustering approach is evaluated. The conclusion of this paper is presented in section 4.

2. DATA AND METHOD

2.1. Datasets

In this study, we used Riau province hotspots data in 2021 as dataset. They take from <http://sipakar.riau.go.id> with total of 3,244 data. For comparison purposes, in this study also used 2 artificial datasets and 7 real life datasets as shown in Figure 1. The first artificial dataset is data points in 2D space (artificial 1 dataset) shown in Figure 1(a). It contains 112 data points divided into 4 clusters and data variance within clusters is 15.9. The second artificial dataset is data points in 3D space (artificial 2 dataset) shown in Figure 1(b). It also contains 112 data points divided into 4 clusters and data variance within clusters is 27.

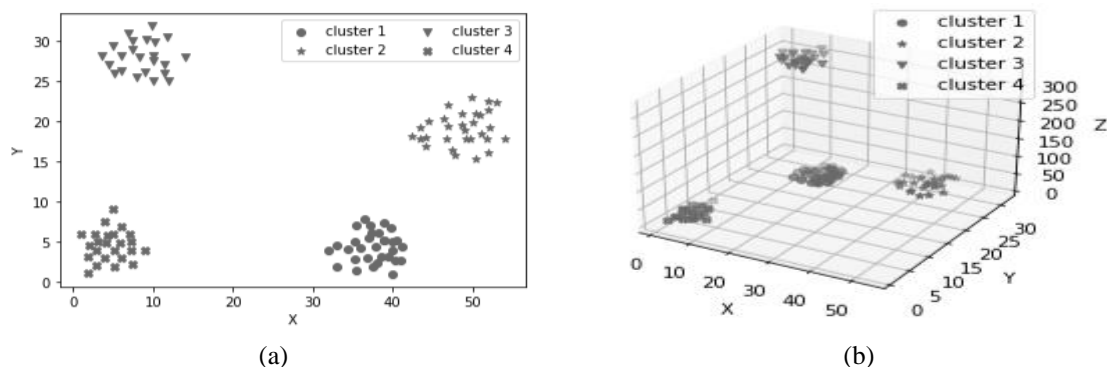


Figure 1. Artificial dataset (a) artificial 1 dataset (b) artificial 2 dataset

The three real-life datasets are from University of California at Irvine (UCI) repository, namely wine dataset, iris dataset and breast cancer dataset. And the others, two real-life datasets are from Kaggle repository, namely crude oil dataset and glass dataset. They were normalized before being clustered. This was to equate the range of parameters for each attribute of them.

2.2. Method

In this section we describe the steps of the GAP K-means method. This method was used as a clustering method in this study and compared with other existing clustering methods. Clustering evaluation that used to measure the clustering performance is explained in this section.

2.2.1. Genetic algorithm with polygamy K-means

GAP K-means is an improvement of K-means. Usually the initial centroid of K-means is determined randomly. In GAP K-means, the initial cluster center's are specified using the GAP method. GAP method is modification of GA, crossover in GAP with polygamy mating. One father married more than one mother. The steps of GAP K-means are follows:

Step 1. The initial parameter setting

Parameter input in GAP K-means is *n*-cluster. In this study, the *n*-cluster variation for clustering of hotspot data was *n*=2,3,4,5 and 6. Meanwhile, *n*-cluster of artificial and real-life datasets are adjusted according to the number of cluster or classes of the dataset. *n*-cluster of artificial 1 and artificial 2 is 4 respectively. The number of classes in the iris dataset, wine dataset and breast cancer dataset are 3,3, and 2 respectively. Input parameters for initialization of the centroid include: number of population (*P*), mutation rate (*MR*), crossover rate (*CR*), max number of iterations and *n*-mating. *n*-mating is the number of mating desired in crossover of GAP method.

Step 2. The initialization of K-means centroid with GAP method

GAP steps are same with GA steps. Crossover process in GAP is modified from crossover in GA. This process to get the best initial centroid for the clustering process using K-means in step 3. Procedure of GAP method is follows:

- a) Initial population selection. Population is formed from chromosomes. Each chromosome is encoded from *k* centroid. Each centroid initialized from dataset randomly. *k* corresponds to the number of clusters. This process is until *P* formed. In this study, *P*=30 chromosomes.
- b) Fitness computation. This step is used to evaluate chromosomes based on fitness function *f(c)*. Fitness function is defined as (1):

$$f(c) = \frac{1}{D} \tag{1}$$

D is computed as follow (2):

$$D = \sum_{i=1}^k \sum_{x_j \in C_i}^n d(x_j, c_i) \tag{2}$$

where *k* denotes the clusters number, $x_j \in C_i$ denotes data in each cluster and $d(x_j, c_i)$ is distance between each data x_j with c_i centroid of the cluster C_i . Calculation of distance in this study using Euclidean distance. Maximization of the fitness function when minimizing *D*

- c) Selection. This step selects the best individual or chromosome based on fitness value as father. *n*-mother chromosomes selected by selection method. In this study we use Roulette Wheel selection. *n*-mother chromosomes correspond to *n*-mating, where *n* > 1
- d) Elitism. In this step, the best individual or chromosome is retained to maintain population quality. This also has an impact on the quality of initial centroid.
- e) Crossover. One selected father is mated with *n*-mother in this step. Each mating with one mother as shown in (3):

$$c_i = \alpha_1 X + \alpha_2 Y \tag{3}$$

where c_i is new centroid (offspring of mating his parents). *X* is father, *Y* is mother, α_1 is random (0,1) and $\alpha_2 = 1 - \alpha_1$. Every mating of two parent will produce two offspring. From *n*-mated increase 2*n* new chromosomes.

- f) Mutation. In this step, individuals can mutate based on the mutation rate. For example, if mutation rate is 0.1. It is expected that 10% of genes in the population will change.

Step 3. K-means clustering

Output from Step 2 as input for clustering using K-means. This method as follow:

- k centroid from the result of initialization centroid using GAP process.
- Distance between each data point x_j and centroid c_i $d(x_j, c_i)$ was calculated using Euclidean distance.
- Set data point into the centroid, whose distance of data point with centroid is the nearest of all centroids.
- Recalculate the centroid k position if all the objects are placed.
- Repeat steps b and c until the centroid k position does not change.

Summary of the algorithm of GAP K-means presented in pseudo code according to Algorithm 1 and 2. Pseudo code of GAP K-means shown in Algorithm 1. Algorithm 2 is pseudo code of determination of initial cluster center using GAP.

Algorithm 1. Genetic algorithm with polygamy

K-means (GAP K-means)

Parameters: nCluster, dataset;

Return: SSE, SC, DBI, Fmeasure;

```

1: Process: initCentroids =
  Genetic Algorithm Polygamy
2: While: stopping criteria
3:   Compute: distance each data;
4:   Determine: cluster each data;
5:   Compute: SSE;
6:   Process: update centroids;
7: Compute: Fmeasure, DBI, SC;

```

Algorithm 2. Genetic algorithm with polygamy

Parameters: Mr, Cr, nPop, maxLoop, nMate;

Return: bestInd;

```

1: Initialization: pop;
2: While: stopping criteria
3:   Compute: fitness function(pop);
4:   Determine: bestInd;
5:   Process: Elitism;
6:   Loop: nMate
7:     Process: Roulette Wheel Selection;
8:     Process: Crossover;
9:     Process: Mutation;

```

2.2.2. Clustering evaluation

Clustering evaluation is measurement of cluster' validity. To compare the clustering method performance, four clustering evaluations are used. They are SSE, SC, DBI, and F-measure.

SSE: SSE is often used as research criteria for determining the optimal cluster [27]. In clustering, SSE can compute as shown in (4).

$$SSE = \sum_{i=1}^k \sum_{j=1}^n_{x_j \in C_i} d(x_j, z_i)^2 \quad (4)$$

SC: SC is one of the intra cluster evaluation. SC can be calculated [5] as shown in (5).

$$SC = \frac{1}{N} \sum_{i=1}^N s_i \quad (5)$$

Where N is number of clusters, s_i is Silhouette index that compute as shown in (6).

$$s_i = \frac{b_i - a_i}{\max\{(a), (b)\}} \quad (6)$$

a_i is mean distance between data and all of data (i) in the same cluster. b_i is mean distance between data (i) and all data in nearest cluster. Silhouette index values range from -1 to 1. If Silhouette index value close to 1, the higher value of the object ownership level in the cluster.

DBI: this index is function of ratio of total *within-cluster* variance to *between-cluster* variance. DBI is define as shown in [28] (7):

$$DBI = \frac{1}{k} \sum_{i=1}^k R_{i,qt} \tag{7}$$

where $R_{i,qt} = \max_{j,j \neq i} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij}} \right\}$. Variance of *within-cluster* i th (S_i) can compute $S_i = \frac{1}{c_i} \sum_{x_j \in C_i} d(x_j, z_i)$. And variance of *between-cluster* (d_{ij}) is $d(z_i, z_j)$, where z_i represent the i th centroid of cluster. Smaller DBI value achieve more accurate clustering.

F-Measure: F value used to find out the prediction of clustering of the algorithm according to the cluster. The value F=1 means that for each object it has already find corresponding cluster. So, in this study the measurement used for clustering of two of artificial datasets and three real-life datasets from UCI. Total F value of the best fitting found cluster as shown in (8)-(10):

$$F = \sum_{c \in C} \frac{|c|}{\sum_{c \in C} |c|} F(c) \tag{8}$$

which,

$$F(c) = \max_{c' \in C'} F(c', c) \tag{9}$$

and,

$$F(c', c) = \frac{rec(c',c)prec(c',c)}{\frac{1}{2}(rec(c',c)+prec(c',c))} \tag{10}$$

each cluster $c' \in C'$ and cluster $c \in C$, recall measure can define $rec(c', c) = |c \cap c'|/|c|$ and precision measure $prec(c', c) = |c \cap c'|/|c'|$ [29].

3. RESULTS AND DISCUSSION

GAP K-means is implemented for clustering on hotspot data. Performance of GAP K-means on this data compare with K-means and GA K-means were shown Tables 1-3. Performance is measured by three evaluation clustering, SSE, SC and DBI. Parameter iteration show the number iteration of K-means in each method. It is important to note that the properties in this case a Dell laptop Computer running on Windows 10, Processor Intel(R), Core (TM) i3-5005U CPU @ 2.00 GHz 2.00 GHz, 8.00 GB RAM, 64-bit operating system, x-64-based processor was used.

Table 1. The performance of K-means and performance of GA K-means, GAP K-means with variance of mutate rate and crossover rate on hotspot data at k cluster=5

MR	EC	CR=0.6		CR=0.7		CR=0.8		K-means
		GAP KM	GA KM	GAP KM	GA KM	GAP KM	GA KM	
0.1	SSE	59.88581	59.88365	59.88365	59.88654	59.88654	59.88495	59.99525
	SC	0.428662	0.428604	0.428604	0.428667	0.428667	0.428596	0.429659
	DBI	0.815833	0.815559	0.815559	0.816173	0.816173	0.815594	0.804134
	time	8	7	8	9	15	10	11
0.15	SSE	59.8861	59.88449	59.88654	59.88654	59.88365	59.88654	
	SC	0.42867	0.42861	0.428667	0.428667	0.428608	0.428667	
	DBI	0.815914	0.815485	0.816173	0.816173	0.815901	0.816173	
	time	9	8	10	12	10	14	
0.2	SSE	59.88613	59.88654	59.96421	59.8856	59.8861	59.9099	
	SC	0.428667	0.428667	0.42897	0.428622	0.42867	0.428837	
	DBI	0.815572	0.816173	0.806475	0.815881	0.815914	0.811251	
	time	9	7	12	14	9	10	
0.25	SSE	59.89455	59.88654	59.88365	59.8842	59.88654	59.8861	
	SC	0.428759	0.428667	0.428608	0.428596	0.428667	0.42867	
	DBI	0.815423	0.816173	0.815901	0.815594	0.816173	0.815914	
	time	8	12	17	8	13	8	
0.3	SSE	59.88654	59.88654	59.88386	59.88521	59.88654	59.8858	
	SC	0.428667	0.428667	0.428608	0.428608	0.428667	0.428647	
	DBI	0.816173	0.816173	0.815901	0.815901	0.816173	0.815942	
	time	11	14	13	20	13	9	

Optimal cluster on hotspot data is known by the elbow method based on K-means. The result k optimal cluster is 5. The n -mating used in GAP K-means is $n=4$. Table 1 shown that generally, performance of GA K-means and GAP K-means are better than K-means based on SSE and time to convergence. Based on SC, GAP K-means performance is better than GA K-means. This shows that the initial centroid of the GAP K-means provides a better form cluster than GA K-means. But based on SSE, DBI and time to convergence, GA K-means is better. The performance of GAP K-means and GA K-means is influenced by the determination of mutation rate and crossover rate.

The initial centroid location of each method at k cluster=5 with GA K-means and GAP K-means using Mutation rate=0.1 and Crossover rate=0.7 are shown at Figure 2. From Figure 2, it can be seen that both GA K-means and GAP K-means methods provide initial centroid almost representing the final result of clustering compared K-means. Initial centroids of both GA K-means and GAP K-means also almost close together. This is what causes both methods to converge faster than K-means.

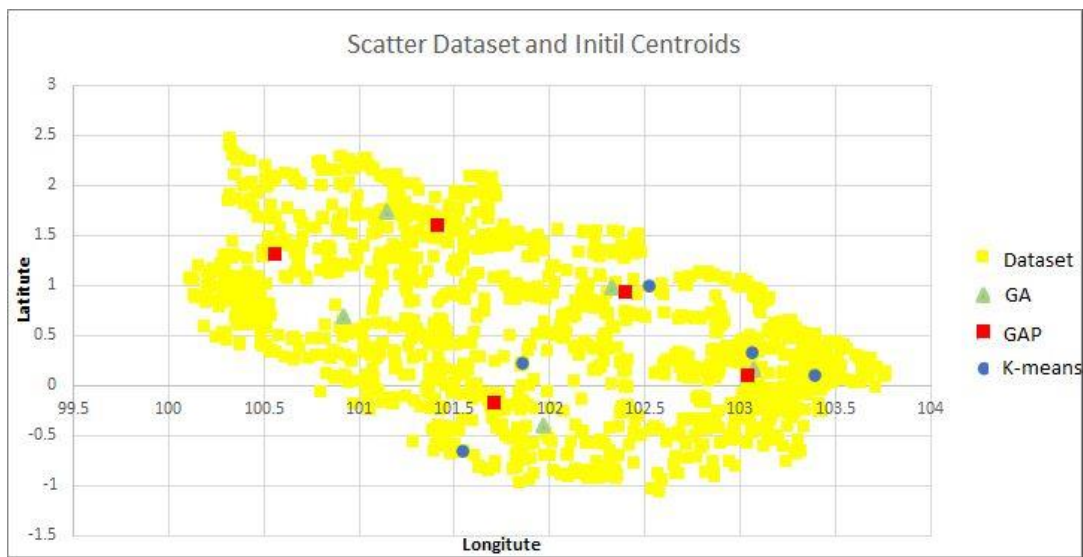


Figure 2. Initial centroid of K-means, GA K-means, and GAP K-means

Table 2 shown the performance of K-means, GA K-means and GAP K-means based SSE, SC, DBI, and F-measurement on two artificial datasets. Based on SSE, GA K-means outperform to others. Time to convergence of GA K-means is also faster than GAP K-means. DBI of GAP K-means is best than others. In artificial 1 dataset, GAP K-means outperform to others based on SC, DBI, and F value. Based on the SC value, indicating the initial centroid of the GAP K-means method gives a better form of cluster than both K-means and GA K-means.

Table 2. The performance of K-means, GA K-means, and GAP K-means on artificial 1 dataset, and artificial 2 dataset

Performance	Artificial 1			Artificial 2		
	K-means	GA-K	GAP-K	K-means	GA-K	GAP-K
SSE	4.906481	1.649987	2.120576	6.546336	2.006083	2.034144
SC	0.560289	0.732878	0.779182	0.597405	0.797996	0.763517
DBI	0.717985	0.39397	0.317732	0.745351	0.295117	0.348619
F	0.786806	0.948328	0.973383	0.784094	0.973887	0.959095
time	4	3.066667	3.333333	6	3.2	3.75

Tables 3 and 4 shown the performance of K-means, GA K-means and GAP K-means based SSE, SC, DBI, and F-Measurement on real-life datasets. From Table 3 especially iris dataset, GAP K-means and GA K-means are better than K-means. On breast cancer dataset, all three methods show the same performance value but GAP K-means was faster in achieving convergent than other methods. From Table 4, that generally we know, GAP K-means was faster than GA K-means and K-means in achieving convergent, except on glass dataset. Based on SSE, GAP

K-means outperform than others. Like breast cancer dataset, crude oil dataset shows the same performance value and time to converge of GAP K-means was faster than others. GAP K-means also faster to convergence than K-means. And based SSE value, GAP K-means outperform than others on glass dataset.

Table 3. The performance of K-means, GA K-means, and GAP K-means on iris dataset, wine dataset and breast cancer dataset

Performance	Iris			Wine			Breast cancer		
	K-means	GA-K	GAP-K	K-means	GA-K	GAP-K	K-means	GA-K	GAP-K
SSE	7.12275	7.032184	7.060195	48.96052	48.97116	48.97275	215.8383	215.8383	215.8383
SC	0.482929	0.493121	0.488753	0.300894	0.300004	0.29979	0.384549	0.384549	0.384549
DBI	0.786733	0.774387	0.779678	1.30864	1.311123	1.314802	1.136335	1.136335	1.136335
F	0.714458	0.72727	0.721779	0.84168	0.829556	0.831111	0.623086	0.623086	0.623086
time	6	6.933333	9.133333	3	7.666667	6.6	6	7.6	6.933333

Table 4. The performance of K-means, GA K-means, and GAP K-means on crude oil price dataset and glass dataset

Performance	Crude oil price			Glass dataset		
	K-means	GA-K	GAP-K	K-means	GA-K	GAP-K
SSE	7.978462	7.978462	7.978462	34.13721	34.13663	34.13606
SC	0.609172	0.609172	0.609172	0.526283	0.525052	0.523822
DBI	0.562868	0.562868	0.562868	1.053373	1.055519	1.057665
F-value	0.002633	0.002633	0.002633	0.312451	0.309187	0.305923
time	8	6	5.9	3	4.2	5.2

4. CONCLUSION

GAP K-means is development of the GA method which was previously able to overcome problem in K-means related to the determination of initial centroids. The polygamy crossover is expected to speed up process of finding the optimal initial centroid. So, GAP K-means is expected to have better performance with time to converge faster than GA K-means. Based on this experimental, GAP K-means were able to overcome the problem of K-means related to initial centroid. This were shown to perform better than K-means. But unfortunately, it has not been able to overcome long of time GA K-means reaches convergence.

ACKNOWLEDGEMENTS

This research was supported by DTRPM Direktorat Riset, Teknologi dan Pengabdian Masyarakat by the scheme of "Penelitian Kerja Sama Perguruan Tinggi" No: 04/KP/LPPM/ITATS/2022 and BPPD Riau.




REFERENCES

- [1] M. Zainal, U. Suworo, D. Mariana, and S. I. Redjo, "Governance of forest and peatland fire prevention in Riau Province," *In International Conference on Democracy, Accountability and Governance (ICODAG 2017)*, vol. 163, pp. 122–125, 2017, doi: 10.2991/icodag-17.2017.23.
- [2] "Walhi: Natural Factors Have More Influence on Reducing Karhutla Haze," CNN Indonesia, 2021, Accessed: Nov. 16, 2022. [Online]. Available: <https://www.cnnindonesia.com/nasional/20211025175130-20-712148/walhi-faktor-alam-lebih-berpengaruh-tekan-kabut-asap-karhutla>.
- [3] H. L. Tata, B. H. Narendra, and Mawazin, "Forest and land fires in pelalawan district, Riau, Indonesia: Drivers, pressures, impacts and responses," *Biodiversitas*, vol. 19, no. 2, pp. 494–501, 2018, doi: 10.13057/biodiv/d190224.
- [4] H. A. Adrianto, D. V. Spracklen, S. R. Arnold, I. S. Sitanggang, and L. Syaufina, "Forest and land fires are mainly associated with deforestation in Riau Province, Indonesia," *Remote Sensor*, vol. 12, no. 1, pp. 1–12, 2020, doi: 10.3390/RS12010003.
- [5] R. Trisminingsih and S. S. Shaztika, "ST-DBSCAN clustering module in SpagoBI for hotspots distribution in Indonesia," *In 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pp. 327–330, 2017, doi: 10.1109/ICITACEE.2016.7892465.
- [6] A. Hatamlou, "Black hole: A new heuristic optimization approach for data clustering," *Information Sciences. (Ny)*, vol. 222, pp. 175–184, 2013, doi: 10.1016/j.ins.2012.08.023.
- [7] S. Louhichi, M. Gzara, and H. Ben Abdallah, "A density based algorithm for discovering clusters with varied density," *In 2014 World Congress on Computer Applications and Information Systems (WCCAIS)* no. 1, 2014, doi: 10.1109/WCCAIS.2014.6916622.
- [8] M. Zhou, H. Huang, and Q. Wang, "A graph-based clustering algorithm for anomaly intrusion detection," *In 2012 7th International Conference on Computer Science and Education (ICCSE)*, pp. 1311–1314, 2012, doi: 10.1109/ICCSE.2012.6295306.
- [9] G. Karypis, E.-H. (Sam) Han, and V. Kumar, "CHAMELEON: A hierarchical clustering algorithm using dynamic modeling," *IEEE Computer*, vol. 32, no. 8, pp. 68–75, 1999, doi: 10.1109/2.781637.
- [10] J. Hämäläinen, S. Jauhiainen, and T. Kärkkäinen, "Comparison of internal clustering validation indices for prototype-based clustering," *Algorithms*, vol. 10, no. 3, 2017, doi: 10.3390/a10030105.
- [11] S. Merendino and M. E. Celebi, "A simulated annealing clustering algorithm based on center perturbation using Gaussian mutation," *In The Twenty-Sixth International FLAIRS Conference*, no. 2, 2013, pp. 456–461.




- [12] S. Askari, "Fuzzy C-means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: review and development," *Expert System Application*, vol. 165, p. 113856, 2021, doi: 10.1016/j.eswa.2020.113856.
- [13] R. R. Muhima, M. Kurniawan, and O. T. Pambudi, "A LOF K-means clustering on hotspot data," *International Journal of Artificial Intelligence and Robotics*, vol. 2, no. 1, pp. 29–33, 2020, doi: 10.25139/ijair.v2i1.2634.
- [14] M. A. Rahman and M. Z. Islam, "A hybrid clustering technique combining a novel genetic algorithm with K-means," *Knowledge-Based System*, vol. 71, no. August, pp. 345–365, 2014, doi: 10.1016/j.knosys.2014.08.011.
- [15] A. Jahwar, "Meta-heuristic algorithms for K-means clustering: a review," *PalArch's Journal of Archaeology of Egypt/Egyptology*, vol. 17, no. 7, pp. 7–9, 2021.
- [16] A. E. Ezugwu, A. K. Shukla, M. B. Agbaje, O. N. Oyelade, A. José-García, and J. O. Agushaka, "Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature," *Neural Computing and Applications*, vol. 33, no. 11, 2021, doi: 10.1007/s00521-020-05395-4.
- [17] M. S. Mahmud, M. M. Rahman, and M. N. Akhtar, "Improvement of K-means clustering algorithm with better initial centroids based on weighted average," in *2012 7th International Conference on Electrical and Computer Engineering*, pp. 647–650, 2012, doi: 10.1109/ICECE.2012.6471633.
- [18] L. Li, X. Zhou, Y. Li, J. Gu, and S. Shen, "An improved genetic algorithm with lagrange and density method for clustering," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 24, pp. 1–14, 2020, doi: 10.1002/cpe.5969.
- [19] H. Xie *et al.*, "Improving K-means clustering with enhanced firefly algorithms," *Applied Soft Computing*, vol. 84, p. 105763, 2019, doi: 10.1016/j.asoc.2019.105763.
- [20] Y. Li, X. Zhou, J. Gu, K. Guo, and W. Deng, "A novel K-means clustering method for locating urban hotspots based on hybrid heuristic initialization," *Application Science*, vol. 12, no. 16, 2022, doi: 10.3390/app12168047.
- [21] T. Niknam and B. Amiri, "An efficient hybrid approach based on PSO, ACO and K-means for cluster analysis," *Application Soft Computer Journal*, vol. 10, no. 1, pp. 183–197, 2010, doi: 10.1016/j.asoc.2009.07.001.
- [22] X. Yin, "Construction of student information management system based on data mining and clustering algorithm," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/4447045.
- [23] U. Maulik and S. Bandyopadhyay, "Genetic algorithm-based clustering technique," *Pattern Recognition*, vol. 33, 2000, doi: 10.1016/S0031-3203(99)00137-5.
- [24] A. M. Aibinu, H. Bello Salau, N. A. Rahman, M. N. Nwohu, and C. M. Akachukwu, "A novel clustering based genetic algorithm for route optimization," *Engineering Science and Technology, an International Journal*, vol. 19, no. 4, pp. 2022–2034, 2016, doi: 10.1016/j.jestch.2016.08.003.
- [25] A. M. Aibinu, H. B. Salau, C. M. Akachukwu, and M. N. Nwohu, "Polygamy based genetic algorithm for unmanned aerial vehicle (UAV) power optimization: A proposal," in *2014 11th International Conference on Electronics, Computer and Computation (ICECCO)*, 2014, doi: 10.1109/ICECCO.2014.6997555.
- [26] R. R. Muhima, M. Kurniawan, S. R. Wardhana, and A. Yudhana, "N-mating effect on genetic algorithm-based clustering performance for hotspots data," in *2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, 2022, pp. 212–215. doi: 10.1109/COMNETSAT56033.2022.9994400.
- [27] R. Nainggolan, R. Perangin-Angin, E. Simarmata, and A. F. Tarigan, "Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the elbow method," in *Journal of Physics: Conference Series*, vol. 1361, no. 1, 2019, doi: 10.1088/1742-6596/1361/1/012015.
- [28] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002, doi: 10.1109/TPAMI.2002.1114856.
- [29] C. Wartena and R. Brussee, "Topic detection by clustering keywords," in *2008 19th International Workshop on Database and Expert Systems Applications*, pp. 54–58, 2008, doi: 10.1109/DEXA.2008.120.

BIOGRAPHIES OF AUTHORS






Rani Rotul Muhima    is lecturer at Department of Informatics Engineering, Institut Teknologi Adhi Tama Surabaya, Indonesia. She received the B.Sc. degree in Physics from Airlangga University, Surabaya, Indonesia. M.Eng. degree in Electrical Engineering with specialization in game technology from Institut of Sepuluh Nopember, Surabaya, Indonesia. Her research areas are data mining, research interests include data mining, artificial intelligence, and game technology. She can be contacted at email: rani.muhima@itats.ac.id.






Muchamad Kurniawan    is lecturer at Department of Informatics Engineering, Institut Teknologi Adhi Tama Surabaya, Indonesia. He received the B.Sc. degree in Informatics Engineering from Institut Teknologi Adhi Tama Surabaya, Indonesia. M.Eng. degree in Informatics Engineering from Institut of Sepuluh Nopember, Surabaya, Indonesia. His research areas are data mining, machine learning and optimization. He can be contacted at email: muchamad.kurniawan@itats.ac.id.






Septiyawan Rosetya Wardhana    is lecturer at Department of Informatics Engineering, Institut Teknologi Adhi Tama Surabaya, Indonesia. He received the B.Sc. degree in Informatics Engineering from Institut Teknologi Adhi Tama Surabaya, Indonesia. M.Eng. degree in Informatics Engineering from Institut of Sepuluh Nopember, Surabaya, Indonesia. His research areas are machine learning and natural language processing. He can be contacted at email: rossywardhana@gmail.com.






Anton Yudhana    is lecturer at Department Electrical Engineering Ahmad Dahlan University. He received the B.Sc. degree in Electrical Engineering from Institut of Sepuluh Nopember, Surabaya, Indonesia. M.Eng. degree in Electrical Engineering from Gadjah Mada University, Jogjakarta, Indonesia and Ph.D. degree in Electrical Engineering. His research now focuses on automation and communication applications in agriculture and health. He is director of research and community service Ahmad Dahlan University. He is inventor and founder AgriPrecision product (Simonkori, PuDang, and E-Kalpin). He received an award as a lecturer at the faculty of industrial technology, Ahmad Dahlan University with the best publication in the year 2020. He is keynote speaker at various international conference. He can be contacted at email: eyudhana@ee.uad.ac.id.



Sunardi    received the B.Sc. degree from Gadjah Mada University. M.Eng. degree from Institut Teknologi Bandung and Ph.D. degree from University Technology Malaysia. He is the head of the Faculty of Industrial Technology, Ahmad Dahlan University. His research areas are geographical information on system, forensic analysis, steganography, and internet of things. He can be contacted at email: sunardi@mti.uad.ac.id.



Mitra Adhimukti    is the head of prevention division of Regional Disaster Management Agency of Riau Province, Pekanbaru Baru, Indonesia. He received the B.Sc. degree in Geological Engineering from Padjadjaran University. M.Eng. degree in Information System Management from Gunadarma University. He can be contacted at email: adhimukti@yahoo.com.