Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11598

e-ISSN : 2541-2019
p-ISSN : 2541-044X

OPTIMIZATION MODEL IN CLUSTERING THE HAZARD ZONE AFTER AN EARTHQUAKE DISASTER

Monica Natalia br Bangun[1,] Open Darnius[2,] Sutarman [3]
University North Sumatera, Medan, Indonesia
[1)] monicanatab@gmail.com, [2)] opendarnius@gmail.com, [3)] sutarman@usu.ac.id

Abstract. There are a large number of approaches to clustering problems, including optimization-based methods involving mathematical programming models to develop efficient and meaningful clustering schemes. Clustering is one of the data labeling techniques. K-means clustering is a partition clustering algorithm that starts by selecting k representative points as the initial centroid. Each point is then assigned to the nearest centroid based on the selected specific proximity measure. This writing is focused on the grouping of post-earthquake hazard zones based on grouping with regard to certain characteristics which aim to describe the process of partitioning the N-dimensional population into K-sets based on the sample. This research consists of three steps, namely standardization, data clustering using K-means and data interpolation using the K-means clustering algorithm and zoning of 7 variables, namely magnitude, depth, victim died, the victim didn't die, public facilities were heavily damage, public facilities were slightly damage, and affected areas.

Keywords: Earthquake, Hazard, K-means clustering

INTRODUCTION

Indonesia is located at the confluence of active tectonic plates, active mountain paths, and tropical climates, making some of its areas vulnerable to natural disasters. The most frequent natural disasters are earthquakes. This is because it occurs at shallow depths and has a large enough magnitude and is located near settlements and population activities. With the study of the hazard of ground shaking, there is a basis for making regional spatial planning policies based on earthquake hazard mitigation. The first step that can be taken is to determine the hazard zone area to facilitate the evacuation of people affected by the earthquake. Utilization of earthquakes can be done by grouping the data according to the information in the data, so that the hazard zones after the earthquake can be known.

According to (Senduk et al., 2019), grouping is an unsupervised learning, where a group of data is directly grouped based on the level of similarity without supervision. Each group, called a cluster, consists of objects that are grouped based on the principle of maximizing intraclass similarity and minimizing interclass similarity. That is, object clusters are formed so that objects in the cluster have a high similarity compared to each other, but are somewhat different from objects in other clusters (Sun et al., 2012).

Clustering algorithm has been applied to a variety of problems, including exploratory data analysis, data mining, image segmentation and mathematical programming. K-means is one of the general methods for partitioning which is quite efficient in terms of variance in groups. K-means clustering groups data

*Corresponding author

groups into a predetermined number of clusters, based on the Euclidean distance as a measure of similarity. The purpose of the K-Means method is to minimize data variation in the same cluster while in different clusters the variation in data will be maximized (Witten et al., 2011). K-means clustering is the most widely used partition clustering algorithm, and one of the simplest and most efficient clustering algorithms proposed in the data clustering literature. The K-means procedure is easy to program and computationally economical, making it feasible to process very large samples on a digital computer. The concept of K-means represents a generalization of the average of ordinary samples and is naturally geared towards studying the asymptotic behavior in question, the object of which is to establish some kind of law of large numbers for K-means.

Decision-making problems are often formulated as optimization problems. Mathematical optimization will model various problem cases and find the right and fast way or method to solve it. Mathematical optimization is aimed at methods to obtain a solution that maximizes an objective function and minimizes risk. Based on the evidence above, this study will cluster all earthquake events that occurred in Indonesia for 5 years to see the patterns that occur, making it easier to classify the hazard zone areas after the earthquake.

LITERATURE REVIEW

Data Clustering

Data grouping is one of the data labeling techniques (Aggarwal & Reddy, 2014). In data grouping, given unlabeled data and must put similar samples in one pile, called clusters, and different samples must be in different clusters. Clustering is useful in several machine learning and data mining tasks including image segmentation, information retrieval, pattern recognition, pattern classification, network analysis, and so on. This can be seen as an exploratory task or a preprocessing step. If the goal is to explore and reveal hidden patterns in the data, clustering becomes an exploratory task in its own right. However, if the resulting cluster will be used to facilitate other data mining or machine learning tasks

Clustering Methods

The clustering methods (Gulia, 2016) can be classified into the following categories:

- **Partitioning Method**

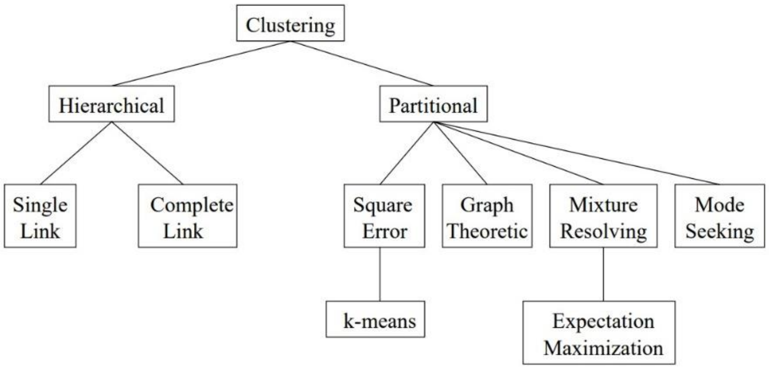
  Suppose given object database 'n' and partition method construct partition 'k' data. Each partition will represent a cluster and k n. This means it will classify the data into k groups, which satisfies the requirement that each group contains at least one object and each object must belong to exactly one group. For a specified number of partitions (e.g., k), the partition method creates the initial partition. Then use iterative relocation techniques to increase the partition by moving objects from one group to another.
- **Hierarchical Method**

  This method creates a hierarchical decomposition of a given set of data objects. There are two approaches here. First, the Agglomerative (bottom up) approach starts with each object forming a separate group. It keeps merging objects or groups that are close to each other. It continues to do so until all groups are merged into one or until the termination condition applies. Second, the Divisional (top-down) approach starts with all objects in the same cluster. In continuous iteration, a cluster is split into smaller clusters. It goes down until every object in a cluster or termination condition applies. This method is rigid, that is, once a merge or split is performed, it can never be undone.

*Corresponding author

Clustering Techniques

The different approaches to data clustering can be explained by the taxonometric representation of the clustering methodology (Jain et al., 1999). There is a difference between hierarchical and partitional approaches (the hierarchical method returns a series of nested partitions, whereas the partitioning method returns only one) (Madhulatha, 2012).



The partitioning method has advantages in applications involving data sets. Partitioning techniques typically generate clusters by optimizing a criterion function defined either locally (on a subset of patterns) or globally (defined across all patterns). The combinatorial search of the set of possible labels for the optimum value of a criterion is obviously very difficult computationally. Therefore, in practice, the algorithm is usually executed several times with different initial states, and the best configuration obtained from all processes is used as the output cluster.

The most intuitive and frequently used criterion function in partition clustering techniques is the Squared Error Algorithms, which tends to work well with isolated and compact clusters. The squared error for grouping L of patterns of H (containing k clusters) is

$$e^2(\mathscr{X}, \mathscr{L}) = \sum_{j=1}^{K}\sum_{i=1}^{n_j}\|\mathbf{x}_i^{(j)} - \mathbf{c}_j\|^2,$$

Where $x_i{}^{(j)}$is the pattern to belonging to the cluster and is the center of the cluster$i^{th} j^{th} c_j j^{th}$

K-means is the simplest and most commonly used algorithm using the squared error criterion (MacQueen, 1967). It starts with a random initial partition and continues to reassign patterns to clusters based on the similarity between the pattern and the cluster center until the convergence criteria are met (e.g., there is no reassignment of any patterns from one cluster to another, or the squared error stops significantly reducing after several iteration). The K-means algorithm is popular because it is easy to implement, and the time complexity is O(n), where n is the number of patterns. The main problem with this algorithm is that it is sensitive to the initial partition selection and may converge to the local minimum of the criterion function value if the initial partition is not selected correctly (Ahmed et al., 2020).

K-MEANS

K-means clustering is the most widely used partition clustering algorithm. It starts by selecting k representative points as the initial centroid. Each point is then assigned to the nearest centroid based on the selected specific proximity measure (Nagari & Inayati, 2020). The first iteration initializes three random points as centroids. In the next iteration the centroid changes position until it converges. Various measures of proximity can be used in the K-means algorithm when calculating the nearest centroid. The choice can significantly affect the centroid assignment and the quality of the final solution. Various types of measures that can be used here are Manhattan distance (L1 norm), Euclidean distance (L2 norm). The objective function used by K-means is called Sum of Squared Errors (SSE) or Residual Sum of Squares (RSS). Given a dataset *D= {x1, x2, …, xN*} consist of N points, denoted using K-means clustering by *C= {C1, C2, …, Ck …, CK*}. The goal is to find the grouping that minimizes the SSE score. Iterative assignment and update steps of the K-means algorithm aim to minimize the SSE score for a given set of centroids.

$$SSE(C) = \sum_{k=1}^{K} \sum_{x_i \epsilon C_k} ||x_i - C_k||^2$$

$$c_k = \frac{\sum_{x_i \epsilon C_k} x_i}{|C_k|}$$

The steps in the K-means clustering algorithm are:

1) Determine the number of clusters
2) Determine the centroid value

   In determining the value of the centroid for the beginning of the iteration, the initial value of the centroid is done randomly. Meanwhile, if determining the value of the centroid which is the stage of the iteration, the following formula is used:

   $$\bar{v}_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj},$$

   where:
   $v_{ij}$is the centroid/average of the i-th cluster for the j-th variable
   $N_i$is the amount of data that is a member of the i-th cluster
   i, k is the index of the cluster
   $j$is the index of the variable
   x_{kj} is the value of the k-th data in the cluster for the j-th variable
3) Calculates the distance between the centroid point and the point of each object. To calculate the distance can use the Euclidean Distance, namely

   $$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2},$$

   where:
   $D_e$is Euclidean Distance
   $i$is the number of objects,
   $(x, y)$are the coordinates of the object and
   $(s, t)$are the coordinates of the centroid.
4) Object grouping

   To determine cluster members is to take into account the minimum distance of the object. The value obtained in the data membership in the distance matrix is 0 or 1, where the value is 1 for data allocated to clusters and 0 for data allocated to other clusters.
5) Return to stage 2, repeat until the resulting centroid value remains and the cluster members do not move to another cluster.

RESULT AND DISCUSSION

In this study, data were obtained from the BMKG in 2014-2018 with a total of 57 data used as shown in table 1.

Table1. Earthquake Data

No.	Region	Date	Time	X1	X2	X3	X4	X5	X6	X7	X8

Obs			**(WIB)**								
1	Kebumen, Central Java	25-Jan-14	05:14:20	6.5	48	0	0	3	177	4	14
2	South OKU, South Sumatra	31-Mar-14	04:13:42	5.4	10	0	0	1	8	1	1
3	Ambon, Maluku	02-May-14	08:43:34	5.7	10	0	3	0	67	1	3
4	Tanah Datar, West Sumatra	10-Sep-14	17:46:19	5.0	10	0	2	0	238	2	4
5	Ternate, North Maluku	15-Nov-14	02:31:44	7.3	10	0	0	0	0	1	1
…	…	…	…	…	…	…	…	…	….	…	…
…	…	…	…	…	…	…	…	…	…	…	…
55	Lombok, NTB	6-Dec-18	08:02:46	5.4	11	0	0	0	3	1	12
56	Wajo	16-Dec-18	22:06:46	4.4	2	0	0	0	2	1	1
57	Manokwari, West Papua	28-Dec-18	10:03:33	6.1	26	0	0	0	1	1	3

Where :
X1 : Magnitude (SR)
X2 : Depth (Km)
X3 : Victim Dies
X4 : Victim Not Died
X5 : Public Facilities Heavy Damage
X6 : Minor Damage Public Facilities
X7 : Affected Area
X8 : Less Affected Area

Clustering Simulation

The results and discussion of the hazard zone grouping are described starting with data collection, data standardization, correlation test, principal component analysis is carried out if the data used are correlated, then continued with cluster analysis (clusters) and then completed with K-means cluster analysis using the Minitab application.

Before performing the K-means cluster analysis, a new column is added, named Initial. Initial column taken from the Earthquake Depth Scale in the catalog of destructive earthquakes is used as a benchmark in the formation of clusters so that in this grouping it is divided into 3 clusters.
Cluster 1 : Shallow Earthquake (depth < 60km) causes damage big
Cluster 2 : Medium Earthquake (60km <depth> 300km) minor damage
Cluster 3 : Shallow Earthquakes (depth > 300km) are not dangerous
For cluster 1 is given the number 1, cluster 2 is given the number 2, cluster 3 is given the number 3

Next, the K-Means Cluster Analysis was carried out with the initial partition column and the results were named Experiment 1 (Exp 1) as follows:

TABLE 1. FINAL PARTITION EXP 1

	Number of observations	**Within cluster sum of squares**	**Average distance from centroid**	**Maximum distance from centroid**
Cluster1	46	333,647	1,812	11.177
Cluster2	11	52,319	2,069	3,146

As a comparison material for selecting the right cluster, the K-Means Cluster Analysis was carried out again without an initial partition column, and the results were named Experiment 2 (Exp 2) as follows:

*Corresponding author

TABLE 1. FINAL PARTITION EXP 1

	Number of observations	**Within cluster sum of squares**	**Average distance from centroid**	**Maximum distance from centroid**
Cluster1	6	174.950	5,105	8,221
Cluster2	30	26,174	0.838	2.108
Cluster3	21	71,841	1,748	2,681

Based on the results of the sum of cluster 1, cluster 2, and cluster 3 in the column within cluster sum squares, a total of 386,966 for Exp 1 and 272,965 for Exp 2.

So that the cluster used uses the final partition in experiment 2. There are 6 areas in cluster 1 with a large level of damage (danger), 30 areas are in cluster 2 with light damage (less dangerous) and 21 areas are in cluster 3 which is an area not harmful.

Principal Component Analysis (PCA) was pioneered by Karl Pearson in 1901 for nonstochastic variables, PCA is a technique for forming new variables which are linear combinations of the original variables. According to (Jain et al., 1999) PCA is a technique used to simplify data by transforming the data linearly to form a new coordinate system with maximum variance. PCA concentrates on explaining the structure of variance and covariance through a linear combination of the original variables, with the main objective of reducing data and making interpretations. I start with the data on the p variable of the number of n data. As shown in the table, the linear combination of the variables, the main components are obtained, namely:$X_1, X_2, \dots, X_p$

$$PC_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1p}X_p$$
$$PC_2 = \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2p}X_p$$
$$PC_n = \alpha_{n1}X_1 + \alpha_{n2}X_2 + \cdots + \alpha_{np}X_p$$

Based on the results of the centroid cluster (attachment 10), then a clustering model is made using the main components obtained:

For Cluster 1:

$$PC_1 = 4{,}0390X_1 - 0{,}9781X_2 + 1{,}1718X_3 - 0{,}2586X_4 + 0{,}2340X_5 + 0{,}2964X_6 - 0{,}1384X_7 + 0{,}0221X_8$$

For Cluster 2:

$$PC_2 = -\,0{,}7088X_1 + 0{,}7831X_2 + 0{,}2133X_3 - 0{,}0130X_4 + 0{,}1495X_5 + 0{,}0379X_6 - 0{,}0352X_7 + 0{,}0062X_8$$

For Cluster 3:

$$PC_3 = -0{,}1414X_1 - 0{,}8393X_2 - 0{,}6395X_3 - 0{,}0924X_4 - 0{,}2804X_5 - 0{,}1388X_6 + 0{,}0899X_7 - 0{,}0151X_8$$

The first cluster is dominated by the following variables:

$X_1\ (Magnitude) X_3 (Victim\ died) X_5\ (Public\ Facilities\ Heavy\ Damage), X_6\ (Minor\ Damage\ Public\ Facilities)$
$X_8\ (Less\ Affected\ Area)$

The second cluster is dominated by the following variables:

$X_2\ (Depth) X_3 (Victim\ Died) X_5\ (Public\ Facilities\ Heavy\ Damage), X_6\ (\ Minor\ Damage\ Public\ Facilities)$
$X_8\ (Less\ Affected\ Area)$

The third cluster is dominated by the following variables:

$X_7\ (Affected\ Area)$

CONCLUSION

Based on the research that has been done using the K-Means algorithm, it can be concluded that the results of Cluster 1 are 6 areas with major damage (danger), Cluster 2 is 30 areas with light damage (less

*Corresponding author

dangerous) and Cluster 3 is 21 areas which are harmless area. Testing data on Minitab using the K-Means algorithm can display the same 3 (three) classes with manual calculations. So that the K-Means algorithm can be used for clustering the hazard zone after an earthquake.

ACKNOWLEDGMENTS

Would like to acknowledge Mr. Open Darnius and Mr. Sutarman who became my supervisor as well as assisting in the completion of master's studies at the mat5hematics department of University North Sumatera.

REFERENCES

Aggarwal, C. C. & Reddy, C. K. (2014). Data clustering. *Algorithms and Applications. Chapman&Hall/CRC Data Mining and Knowledge Discovery Series, Londra*.

Ahmed, M., Seraj, R. & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.

Gulia, P. (2016). Clustering in Big Data: A Review. *International Journal of Computer Applications*, *975*, 8887.

Jain, A. K., Murty, M. N. & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, *31*(3), 264–323.

MacQueen, J. (1967). Classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Probability*, 281–297.

Madhulatha, T. S. (2012). An overview on clustering methods. *ArXiv Preprint ArXiv:1205.1117*.

Nagari, S. S. & Inayati, L. (2020). Implementation of Clustering Using K-Means Method To Determine Nutritional Status. *J. Biometrika Dan Kependud*, *9*(1), 62.

Senduk, F. R., Indwiarti, I. & Nhita, F. (2019). Clustering of earthquake prone areas in indonesia using k-medoids algorithm. *Indonesia Journal on Computing (Indo-JC)*, *4*(3), 65–76.

Sun, Y., Aggarwal, C. C. & Han, J. (2012). Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. *ArXiv Preprint ArXiv:1201.6563*.

Witten, I. H., Frank, E. & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (Third Edit). Morgan Kaufmann.

*Corresponding author