K-Means Clustering Algorithm for Determination of Clustering of Bangkalan Regional Development Potential

A F Doni¹, Y D P Negera², O A H Maria
¹,²Informatic Departement Engineering, University of Trunojoyo Madura, Indonesia
doni.fatah@trunojoyo.ac.id, yudha.putra@trunojoyo.ac.id, okiemaria@gmail.com

Abstract: Regional potential is the ability of an area that can be developed to improve the area according to its regional characteristics. This potential can cover social, economic, infrastructure, natural resources, and human resources and their work. To improve regional development, a strategic policy is needed that leads to improving the economy of a region. In this study the regional grouping of 3 agricultural, plantation, and livestock groups in Bangkalan Regency consisted of 18 sub-districts using the K-Means Clustering method, from which the data was processed preprocessing using Scaling transformations starting from the lower limit search and the upper limit, looking for the Lower Limit and Upper Limit, then grouping using K-Means Clustering. From the research, it was found that testing in the agricultural sector with 3 clusters for the initial cluster center 5,7,15 obtained final clusters according to the highest yield criteria in cluster 1 compared to clusters 2 and 3. For testing in the plantation sector with 3 clusters for the initial cluster center 4,7,16 the final clusters were obtained according to the highest yield criteria in cluster 2 compared to clusters 1, and 3 and for testing the livestock sector with 3 clusters at the initial cluster center 3,11,13 the final clusters were obtained according to the criteria it was known that more height in cluster 2 is compared to cluster 1 and 3.

Keywords: Clustering, K-Means, Regional Potential, Preprocessing, Sector, Scaling.

1. Introduction

The regional potential is an ability within an area that can be developed to improve the area. This potential can cover social, economic, facilities and infrastructure as well as the potential of natural resources, human beings and the results of human labor. The utilization of regional potential must continue to be pursued optimally for regional development. For this reason, it is necessary to have a strategic policy on various regional potentials which leads to an increase in the development of a region, especially in the agricultural, plantation and livestock sectors so that it can improve the economy of a region. By looking at the various types of potential that exist, the policy emphasis is directed at the potential of each region. To overcome this problem, a study is needed that can provide information about the potential of the region's potential, using data mining techniques to group sub-districts in Bangkalan Regency consisting of 18 sub-districts [1][2][3], with several grouping criteria according to their respective sectors using the K-Means Clustering method which will be known in each sub-district which has the same characteristics grouped into one sector of the same potential and sub-districts that have different characteristics grouped into other potential sectors.

The basic principle of clustering is to group objects in a class that has a very similar resemblance to other objects in the same class (similarity), but it is not very similar to other objects in another class (dissimilarity) [4][5][6]. Clustering one of the Data Mining methods that is non-directive (unsupervised), the label of the training data class is unknown, and given a group of measurements, observations, etc., with the aim of dividing the data into clusters or groups based on similar attributes between the data, for example K–Means [7][8]. The purpose of this clustered data is to minimize objective functions set in the clustering process [9], which generally try to minimize variations in a cluster and maximize variations between clusters [10][11]. In this study, the data will be grouped into several clusters using the K-Means method whose function is to partition data into one or more clusters.
that cover 3 different sectors of agriculture, plantation and livestock sectors. The purpose of this study is to design a system that can be used to determine the potential of the region in accordance with its potential with K-Means Clustering so as to facilitate the government in decision making and as a guide in determining local development policies.

2. Methods

A. Overview of the Model

In this study, a system was developed to categorize regional potential into three sectors, namely agriculture, plantation, and livestock based on data obtained from related agencies. The data used include the agricultural sector including land area, planting area, harvest area, yield, productivity, paddy field area, wetland area, dry land area, rainfall, rainfed, irrigation and population, for the plantation sector covering land area, planting area, harvest area, yield, productivity, paddy field area, wetland area, dry land area, rainfall, rainfed, irrigation and population, while for livestock sector includes livestock population, number of livestock owners, meat production, milk production, egg production, population.

After the data is obtained, the data processing (preprocessing) is carried out using several data transformation techniques, which are intended to simplify the data before the grouping process is carried out, after making preprocessing from each data, the next step is to do the process of grouping data using the K-Means method.

B. System Design

The system developed in this study is described in general on the following block diagram:

![Figure 1. Block diagram of the system](image)

In Figure 2 the flow of information from the system begins with processing raw data, the data is processed by transfer scaling by changing the data that was originally random data into group data whose function is to find the lower and upper limits of each data, then transformed with scaling and stored in a table scaling was then summoned and processed by the clustering technique using the K-Means algorithm and the results of the clustering process and the next step of the analysis and evaluation of the results were carried out using cluster analysis and descriptive analysis. In cluster analysis by looking at the total distance between data and the total distance between clusters where Euclidean calculation is used and matching the final value with sector standard values. Whereas for descriptive analysis by observing cluster behavior that has been formed, where each cluster has differences between other clusters.

C. Preprocessing Data

The process of preprocessing data from raw data into data that is ready to use, aims to transform data into certain formats which process is easier and more effective for user needs, with indicators getting more accurate results, reducing computational time for large scale problems, making data values smaller without changing the information they contain. The transformation of the data used in this study are:

1. Centering Techniques can reduce each data with the average of each existing attribute.

Example of transformation using the centering method in column X, then the formula used is as follows:

\[ \tilde{X} = X - \bar{X} \]  

(1)
the result vector after centering, $X$ is the column vector and $\bar{X}$ is the average of the column in question.

2. **Normalization** Divide each data that has been centered with the standard deviation of the attribute in question [20][21], with the formula:

$$\bar{X} = \frac{X - \bar{X}}{\sigma_x}$$  

(2)

3. **Scaling** is the process of changing data so that it is on a certain scale [20][21]. For example, the data is converted in scale (0.1), the Lower Limit (BB) is 0 and the Upper Limit (BA) is 1, for each data in the column operations can be performed.

$$\bar{X} = \frac{X - \text{min}}{X_{\text{max}} - \text{min}}(BA - BB) + BB$$  

(3)

For the maximum column, value is $X_{\text{max}}$ and the minimum value is $X_{\text{min}}$, for each data to a new scale, while for the Lower Limit (BB) and Upper Limit (BA) the value is according to the calculation results.

D. **K-Means Clustering**

K shows the number of clusters. The K value is determined by the user or use [23]. For cases where there are considerations from experts who are competent or expert in their field, the value of k will be easily determined [9]. But it often happens that the present value must be determined by looking at the data (without any consideration from the expert)[24]. The simplicity of K-Means makes this algorithm widely used in various fields and is a clustering method in partitioning that separates data into different groups so that data that has the same characteristics are grouped into one group and data that has different characteristics are grouped into other groups [10].

![Figure 2. K-Means Algorithm](image)

Description: Euclidean Distance Space is this Distance must fulfill the assumption that the observed variables do not correlate and between variables have the same unit. distance measurement is done by calculating the square root of the sum of the squares of the difference of the value of each variable[6][10].

$$D_{xy} = c_{xy} = \sqrt{\sum_{j=1}^{p}(x_{xj} - y_{yj})^2}$$  

(4)

Where: $P$ = Number of variable dimensions of data, $D_{xy} = C_{xy} =$ Distance between Objects $x$ with $y$, $xxj =$ Value of object $x$ in variables $i - j$, $yyj =$ Value of object $y$ in variable $j$

3. **Result**

Before the calculation is done to find the right preprocessing method, it is assumed that the data used from the agricultural sector can be seen in table 1.
Calculate the distance of data with the cluster center using euclidean distance.

14,351.87 7,274.16 7,080.78 395,066.48 172.43

cluster center randomly, select it

The next step is the grouping process with KMeans. After obtaining the results of the reduction in max and min values, which is 15,248.

Table 1. Initial / Raw Data

| No | Kecamatan | A1  | A2  | A3  | A4  | A5  |
|----|-----------|-----|-----|-----|-----|-----|
| 1  | Kamal     | 6356| 3250| 3106| 176,024.94 | 56.67 |
| 2  | Labang    | 3282| 1680| 1602| 108,549.70  | 67.76 |
| 3  | Kwayar    | 6193| 3146| 3047| 172,890.48  | 56.74 |
| 4  | Modung    | 12556| 6388| 6168| 351,633.48  | 57.01 |
| 5  | Blega     | 15248| 7714| 7534| 430,579.64  | 57.15 |
| 6  | Konang    | 10640| 5374| 5266| 262,890.02  | 49.92 |
| 7  | Galis     | 4818| 2420| 2398| 120,696.64  | 50.33 |
| 8  | Tanah Merah | 10380| 5250| 5130| 274,521.44  | 53.51 |
| 9  | Tragah    | 6737| 3404| 3333| 175,596.00  | 52.68 |
| 10 | Socah     | 10877| 5498| 5379| 292,778.82  | 54.43 |
| 11 | Bangkalan | 7185| 3636| 3549| 209,703.68  | 59.09 |
| 12 | Burneh    | 24834| 12570| 12264| 716,901.48  | 58.46 |
| 13 | Arosbaya  | 16214| 8206| 8008| 464,799.66  | 58.04 |
| 14 | Geger     | 18377| 9282| 9095| 522,424.80  | 57.44 |
| 15 | Kokop     | 9299| 4690| 4609| 235,465.06  | 51.09 |
| 16 | Tanjung Bumi | 5813| 2936| 2877| 150,883.62  | 52.44 |
| 17 | Sepulu    | 7763| 3912| 3851| 203,360.36  | 52.81 |
| 18 | Klampis   | 6288| 3178| 3110| 163,756.04  | 52.65 |

Description: In table 1 for A1: Land area / area, A2: planting area, A3: harvest area, A4: production yield, A5: productivity. Using the Scaling transformation by looking for the Lower and Upper Limits, using the frequency distribution and K-Means. To do the scaling process, look for maximum and minimum values, determine the number of intervals. 1. In determining the maximum and minimum values taken from the initial data (data in table 1), reducing the maximum and minimum values, the goal is to get the range of results of the reduction in max and min values, which is 15,248-3,282 = 11,966.

1. In determining the maximum and minimum values taken from the initial data (data in table 1), reducing the maximum and minimum values, the goal is to get the range of results of the reduction in max and min values, which is 15,248-3,282 = 11,966.

A1 = 11,966. A2 = 6,454. A3 = 5,932. A4 = 329,229.94. A5 = 21.21. then determine the number of classes adjusted for the amount of data, namely 3. Finding the width of the interval, obtained from dividing the range value and dividing by the number of classes, which is 11,966 / 3 = 3989. Interval width calculation results: A1 = 3989, A2 = 2152, A3 = 1978, A4 = 109744, A5 = 8. Determine the lower limit, by using a minimum value as the first range, for the second range adding a minimum value with the class width (3,282 + 3989 = 7.271), and for the third range add the second range value with the class width (7.271 + 3989 = 11,260), then get to the fourth range. So that the results of the range calculation are used to determine the lower range value (lower class limit) and upper range, for the lower range using the range value that has been determined as the beginning of the range value. Whereas to find the upper range value that is by using the second range of the lower range value as the first range.

Whereas to find the upper range value that is by using the second range of the lower range value as the first range. So that the results of the upper and lower range values are obtained for A1 to A3 with Range 1 to Range 3. Look for the Lower and Upper Limits of each predetermined range. To search for the Lower Limit, by means of the lower limit, it is reduced by 0.5 while to determine the Upper Limit, the upper limit is added by 0.5. To find the first range in the A1 data, look for the Lower and Upper Limits. BB = (3,282 - 0.5) = 3,281.50. BA = (7271 + 0.5) = 7,271.50. as well as for the calculation of BB and BA on data A1 to A5 the same way.

Determine groups of data, from initial data to each range. In the first data A1 6,356.00 is matched with the first range or the next range, if the value is still in a certain range then it becomes the range group, as well as for the next process. After knowing the group, then calculate the scaling, like the previous test. After obtaining the results of Scaling Hospitality.

The next step is the grouping process with K-Means. Data is grouped into 3 groups. Then specify the cluster center randomly, select it : K1 : 2 = 3,281.50 1,399.61 1,601.50 87,549.20 158.96, K2 : 4= 14,351.87 7,274.16 7,080.78 395,066.48 172.43, K3 : 7 = 3,793.67 1,259.50 1,867.06 98,598.35 147.94.

Calculate the distance of data with the cluster center using euclidean distance.

Table 2. Results of the euclidean distance calculation in the first iteration

| No | sub-district | A1   | A2   | A3   | A4   | A5   | K1  | K2  | K3  |
|----|--------------|------|------|------|------|------|-----|-----|-----|
| 1  | Kamal        | 4.306.51 | 1.923.35 | 2.103.26 | 115,041.37 | 1610 |   * |     |     |
In the table 2, calculate the data in the first iteration to search for averages, and serve as the center of the new cluster. For the first cluster there is 1, that is the second data, the second cluster is 3, the data is 4,5,6. Whereas for the third cluster there are 3, namely data to 1,3,7. Repeat the steps until the data position has not changed. Because in the 2nd and 3rd iteration the cluster position does not change or is the same as the 2nd iteration, then the iteration is stopped and the final result is obtained by 3 clusters.

Figure 3. The results of grouping using K-Means Clustering Bangkalan area potential with 3 sectors of the (1) agricultural sector, (2) plantation sector and (3) livestock sector

In figure 4 is the result of grouping 3 sectors which consist of agriculture sector, there are 3 clusters namely cluster 1, 2 and 3 along with the group, plantation sector there are 3 clusters namely cluster 1,2 and 3 with the group, while livestock sector there are 3 clusters consisting of clusters 1,2 and 3 and their groups.

4. Conclusion

In accordance with the tests that have been done, get the results of testing in the agricultural sector with 3 clusters for the initial cluster center 5,7,15 using sub-district code table 1, repetition of up to 2 iterations, from the highest final cluster results in cluster 1, 4,5,12,13,14 compared to clustering of 2 groups 1,2,3,7,9,11,15,16,17,18 and clustering of 3 groups 6,8,10 but for cluster 2 the average value of wetland area and type of irrigation with irrigation broader than clusters 1 and 3, while cluster 3 has higher productivity than clusters 1 and 2. Testing of plantation sector with initial cluster center 4,7,16 repetitions up to 3 iterations, with the highest results in cluster 1 of group 1,3,4,6,10,13,15,16,18 compared to the clustering of 2 groups 5,7,8,11,12,14,15, and cluster 3 of the group 2,9,17, but for cluster 3 the average value of bulk rain, the type of irrigation with irrigation is wider than clusters 1 and 2, while for clusters er 2 the overall average value is lower than clusters 1 and 3. In the livestock sector, there are 3 clusters at the initial cluster center 3,11,13. There were 2 iterations with the highest scores in cluster 2 of the group 8,11 compared to cluster 1 in the group 1,3,6,10,12,16,18 and cluster 3 in the group 2,4,5,7,9,13,14,15, 17, but for the average value of meat production and milk production the highest value is found in cluster 3, while in cluster 1 the overall average value is lower than clusters 2 and 3. Furthermore, in future research, it is expected that not only grouping data at the level one district but also more than one district, and data can be merged by using several different methods and using different preprocessing techniques.

References

[1] S. of B. Regency, “Bangkalan In Figures 2010,” Manuscr. - Stat. Bangkalan Regency, no. c, p. 303, 2010.
[2] Statistics of Bangkalan Regency, “Bangkalan In Figures 2009,” Manuscr. - Stat. Bangkalan Regency, p. 303, 2009.
[3] A. S. Ahmar, D. Napitupulu, R. Rahim, R. Hidayat, Y. Sonatha, and M. Azmi, “Using K-Means Clustering to Cluster Provinces in Indonesia,” J. Phys. Conf. Ser., vol. 1028, no. 1, pp. 0–6, 2018.
H. H. Ali and L. E. Kadhum, “K-Means Clustering Algorithm Applications in Data Mining and Pattern Recognition,” Int. J. Sci. Res., vol. 6, no. 8, pp. 1577–1584, 2017.

U. R. Raval and C. Jani, “Implementing and Improvisation of K-means Clustering,” Int. J. Comput. Sci. Mob. Comput., vol. 4, no. 11, pp. 72–76, 2015.

K. R. Ahmed and S. Akter, “Analysis of landcover change in southwest Bengal delta due to floods by NDVI, NDWI and K-means cluster with landsat multi-spectral surface reflectance satellite data,” Remote Sens. Appl. Soc. Environ., vol. 8, no. January, pp. 168–181, 2017.

M. S. Amin, Y. K. Chiam, and K. D. Varathan, “Identification of significant features and data mining techniques in predicting heart disease,” Telemat. Informatics, vol. 36, pp. 82–93, 2019.

G. Liu, J. Yang, Y. Hao, and Y. Zhang, “Big data-informed energy efficiency assessment of China industry sectors based on K-means clustering,” J. Clean. Prod., vol. 183, pp. 304–314, 2018.