Application of K-medoids clustering method for grouping corn plants based on productivity, production, and area of land in East Java

D W Wibowo¹, Y Yunshasnawa¹, A Setiawan², E Rohadi¹ and M K Khabibi¹

¹ Informatics Engineering, Politeknik Negeri Malang, Jalan Soekarno-Hatta 9, Malang, Indonesia
² Electrical Engineering, Politeknik Negeri Malang, Jalan Soekarno-Hatta 9, Malang, Indonesia

*gsyn153@gmail.com

Abstract. Corn plants has an important role in compliance national and international food needs after rice and wheat. According to Ir. Pending Dadih Permana, Director General of Food Infrastructure and Facilities (PSP) of the Ministry of Agriculture, East Java is one of province that the biggest corn producers in Indonesia. But it still cannot supply the increasing market needs in Indonesia. Therefore, it requires a grouping corn plants based on productivity, production, and area of land in East Java in order to increase level of procurement of corn plants in East Java. In process of grouping, it use K-Medoids algorithm. K-Medoids algorithm as known as Partitioning Around Medoids (PAM), which is variants of K-Means method. This is based on the use of medoids, not from observation of mean that owned by each cluster, with purpose reduce the sensitivity of partitions due to the extreme values in the dataset. This algorithm is an algorithm that can produce data that is not sensitive to outliers because one of object with some big value. These deviations can occur from data distribution. Based on the result of grouping corn plants based on productivity, production, and area of land which has been done, there are 3 clusters with at most 21 members, and at least 8 members. From this research, it can concluded that K-Medoids algorithm can help determine potential of the location of corn producing plants in East Java.

1. Introduction

The agricultural sector is the most spotlight sector of the government, because of its important role in supporting the economic growth of the society. Food crops are a basic society need. In Indonesia, crops is a due to every person as mentioned in the law of Indonesian Republic number 7, 1996. In Indonesia, crops are benefit products because can fulfil domestic market.

Corn plants is a type of food crop group that is the most available in Indonesia. Corn plants has important role in compliance national and international food needs after rice and wheat. Corn is a common plant that planted in lowland region. For corn plant development, use of superior and high-quality seeds become one of the efforts that continues to be reviewed and disseminated to farmers. Until now, corn plants still a second strategic commodity after rice because in each area, corn still be the second main foot after rice [1].
East Java is the one of the biggest corn producers in Indonesia. The result of productivity, production, and land of area of corn plants in East Java projected by the East Java Provincial Government to increase in 2018. However, increasing of productivity, production, and land of area are not balanced with industrial and household market demand. Increase market demand make rice shelter (Bulog) in East Java confused looking for corn plants that are ready to be sold again. According to head of Divre Bulog of East Java, Muhammad Hasyim said that since 3 months ago, corn stock in Bulog has been sold out. Even now, Bulog of East Java still looking for corn plants in each of area which has reasonable price [2]. The price from corn farmers is higher than the basic production price of industry. So, Bulog make import is an alternative. However, with that alternative, still does not compliance the increasing market needs.

Data mining is a process to find information with identify pattern in dataset. Data mining can implements to any sector like agriculture sector. In agriculture sector, data relating to agriculture such as the amount of productivity, amount of production, and land area can be used to be analysed to find information.

The process of analytical data can through the process of grouping data from data sets using the clustering method. Clustering method is the one of analytical method for data mining. The purpose of this method is for grouping data with same characteristic.

2. Methods
K-Medoids Clustering is one of K-Means variants. K-Medoids used the medoids to find each of cluster member. K-Medoids use partition for clustering a group of n object to be a group of k cluster. This algorithm use random object for represent every each cluster. Clusters are built by calculating the closeness between medoids and non-medoids objects [3].

K-Medoids Clustering have 7 step. This is the step.

- Set value of k (quantity of cluster) object.
- Pick randomize k object from n object as a medoids.
- Pick randomize k object from n object as a non-medoids.
- Calculate the distance from each n object to medoids object with Euclidean Distance with the equation as follows.

\[ d_{(x,y)} = \sqrt{\sum_{i=1}^{n} (x_i - y_j)^2} \]

- Calculate the distance from each n object to non-medoids object with Euclidean Distance.
- Calculate S for find the result. S can find with this equation.
- \[ S = \text{total cost non-medoids} - \text{total cost medoids} \]
- If S > 0, then the clustering process is stopped. But if S < 0, exchange non-medoids with medoids and find new centroid of non-medoids. Repeat step (b) until S < 0.

Silhouette Coefficient used to look at the quality and strength of the cluster, how well an object placed in a cluster. This method is a combination of cohesion and separation methods. The cohesion method works to measure how close the relation between objects in a cluster, and the separation method which serves to measure how far apart a cluster is with another cluster [4]. Silhouette Coefficient calculation is as follows:

1. Calculate the average distance of a cluster member i suppose with all other cluster member that are in one cluster:

\[ \alpha(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i, j) \]

j is the other cluster member in the cluster A and d (i, j) is the distance between the cluster member i to j.

2. Calculate the average distance from the i cluster member with all the cluster member in another cluster, and taken the smallest value.
\[ d(i, C) = \frac{1}{|A|} \sum_{j \in C} d(i, j) \]

by \( d(i, C) \) is the average distance \( i \) cluster member with all the objects in other clusters \( C \) where \( A \neq C \).

\[ b(i) = \min_{C \neq A} d(i, C) \]

3. Silhouette Coefficient value is:

\[ s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \]

The value of the Silhouette Coefficient is located in the range of -1 to 1. If the value is close to 1, the data in one cluster is better. But if the value is close to -1, that is the worse group in clustering [4].

3. Results and discussion

3.1. Designing end product

In the design of this system are running groove design support system as well as the flow calculation methods used. The flow of a general description of the starting of the process of importing the data, and then the data entered into the database. The data can be viewed on the program and the program doing the calculation of grouping process which publishes the result of grouping, the calculation process of testing which release the result of testing, and the program can also display a graph of actual data comparison [5,6].

In this study, this application system was designed by many artefacts on every phases of software development, there were preliminary use case (Figure 1) which provided many functions and block diagrams which were show the flow of process (Figure 2) including the data flow.

![Use Case Diagram](image)

**Figure 1.** Use case diagram.

The menu structure in application data grouping system corn plants in East Java by using K-Medoids Clustering specially designed web-based PHP programming. Use case diagram describing an interaction between one or more actors with the system being designed. The use case diagram illustrated the application page to be accessed by Admin and User, namely: Login, Manage Data Admin and User, Manage Data Corn Plants, Manage Data Location, Watch and Print Data Corn Plants, Clustering and Testing, and Logout. Block diagram describe the flow of the data from raw data into an analysing data. The data can be imported to database with Excel form. Data from database can be viewed on program with website based. The data can be viewed on graph based on location filtering. There were clustering menu with data from database and show the analysed data result. Data from clustering can be tested in testing menu with Silhouette Coefficient.
3.2. Clustering and testing

From the data, it can be clustered into 2 until unlimited. In this research, the author will apply 2 cluster until 6 clusters. Because the author uses data from 2010 until 2017, so the author clustering the data for each year. In each year, the data clustered into 2 until 6 clusters. After the data entered into cluster, the data can be tested with Silhouette Coefficient and show the good or bad result.

From the results of testing using the Silhouette Coefficient, it was found that there were good and poor quality clusters every year. To produce a good number of clusters each year, it is again grouped based on the results of good cluster quality and the average value of S which approaches 1. The grouping results can be seen in Table 1 and Figure 3.

**Table 1.** The result of silhouette coefficient.

| Year | Cluster | Average S  |
|------|---------|------------|
| 2010 | 3       | 0.316855493|
| 2011 | 2       | 0.241751931|
| 2012 | 2       | 0.397876374|
| 2013 | 2       | 0.006082474|
| 2014 | 2       | 0.337465925|
| 2015 | 2       | 0.079190042|
| 2016 | 2       | 0.187668334|
| 2017 | 2       | 0.255576972|
Figure 3. The result of silhouette coefficient.

The results in the table and graph above show that the average quality of the Silhouette Coefficient value that is good for each year showing a pattern of data with 2 cluster numbers.

4. Conclusions
Based on the Silhouette Coefficient test that was carried out above, it is known from the graph of corn plant grouping data showing a pattern of data with 2 cluster numbers. Corn plants data are tested every year by classifying 2 to 6 clusters and testing with Silhouette Coefficient. The test results show that the graph of the average value will drop at the 3rd cluster. While the average value of the test results every year shows above 0 is in the second cluster.

Acknowledgements
The work was supported by Informatics Engineering, State Polytechnic of Malang. Author thanks to website of Ministry of Agriculture for the data.

References
[1] Amin M and Zaenaty 2012 Respon Petani Terhadap Gelar Teknologi Budidaya Jagung Hibrida Bima 5 di Kabupaten Donggala J. Agrika 61 34–47
[2] Widiarti P 2018 Corn Productivity [Online] Retrieved from: http://surabaya.bisnis.com/read/20181022/532/852019/produksi-jagung-di-jatim-diproyeksikan-65-juta-ton-tahun-ini-. Accessed on 16 November 2018
[3] Kaufman L and Rousseeuw P J 1990 Finding Groups in Data: An Introduction to Cluster Analysis 39 (New York: Wiley) pp 1-38
[4] Pramesi D F, Furqon M T and Candra D 2017 Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot) Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer 19 723-732
[5] Ministry of Agriculture 2017 Agriculture Database [Online] Retrieved from: https://aplikasi2.pertanian.go.id/bdsp Accessed on 16 November 2018

Vercillis C 2009 Business intelligence: data mining and optimization for decision making (Milan: Wiley)