JurnalMantik
Volume 4, Number 2, Aug 2020, pp. 1430-1438 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

# Aplication of Validity Index in K Means and Fuzzy C Means

Jontinus Manullang[1], Pahala Sirait[2], Andri[3]

[123] Jurusan Teknologi Informasi
[123]STMIK Mikroskil, Jl. M.H Thamrin No.140 Kel, Pusat Ps., Kec. Medan Kota, Kota Medan, Sumatera Utara 20212, Indonesia.

E-mail: jhoe6590@gmail.com, pahala@mikroskil.ac.id, andri@mikroskil.ac.id

ARTICLE INFO

Article history:
Received: 12/01/2020
Revised: 22/07/2020
Accepted: 01/08/2020

**Keywords**:
Cluster validity index, Davies bouldin index,kd tree, k means, fuzzy c means

ABSTRACT

K-Means and Fuzzy C-Means Clustering is a method of analyzing data that performs the modeling process without supervision (without supervision) and is a method that groups data by partitioning the system. Clusters Clusters and Fuzzy C-Means will produce different clusters in the same dataset, cluster validity index is a method that can be used to improve the results of clustering generated by the clustering method. This study will use the cluster validity index on the kmeans clustering algorithm and Fuzzy C-Means by calculating the index of validity of each kmeans clustering result with k = 2, ..., kmax (k max determined at the beginning) and the results from Fuzzy C-Means with c = 2, ...., cmax (c max is specified at the beginning). By using the cluster validity index, the most optimal cluster is obtained in the second cluster with the Dbi value = 0.45 in the mean K and the second cluster with the Dbi value = 0.5 in the Fuzzy C Mean, and the results of the clustering are consistent.

Copyright © 2020 Jurnal Mantik.
All rights reserved.

## 1. Introduction

Clustering works by grouping data objects into a certain number of clusters [1] in two ways, namely observations in a cluster that are similar to one another and different from observations in other clusters [2] into a number of clusters according to certain characteristics. The basic concept of grouping by surveying a grouping algorithm is widely known by means of related or comparison[3] There are many grouping algorithms, and it is very difficult to provide a good type of grouping method, but there are several grouping methods consisting of partition methods, hierarchical methods, density-based methods, grid-based methods, model-based methods [4].

In clustering that uses the partition concept, there are three concepts that can be used, namely classical partition, fuzzy partition and probabilistic partition. In classic partition, data is exclusively a member of one cluster. In the fuzzy partition, the membership value of a data in a cluster is located at an interval (0,1). The degree of membership for each data in all clusters is 1. Whereas for the possibility of partitions, the total membership value of data in all clusters does not have to be 1.However, to ensure that the data are members of at least one cluster, the membership value is greater than 0 [5]

The K-means clustering algorithm is sensitive to initial values, random selection of centroid locations at the start of the algorithm, treats variables as unknown numbers and the number of clusters is k, where different initial values may cause differences in the results of K-means algorithm grouping making it possible to select two or more clusters in a cluster [6] A possible solution is to combine K-means with data extraction methods where the results show that the reduction of unsupervised dimensions is closely related to unsupervised learning [7].

The fact that an optimal clustering algorithm does not exist for every case is a major drawback of this technique. [8] It has been established that different initializations of the same algorithm result in different groups in many different environments with different characteristics and no one is best in all cases. Basically, the center of this clustering is still inaccurate [9] and the results for the final cluster are not always guaranteed [10]. Finally, to solve this problem, a heuristic method is proposed to improve the accuracy and efficiency of the k-means cluster algorithm and the fuzzy C-means algorithm. The modified algorithm was then applied for grouping the data [9].

JurnalMantikis licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Kd Tree K - Means Clustering [11] is an optimization algorithm of the algorithm [12] which uses the value of density and distance between points / data in determining the initial position of the midpoint of the k means clustering cluster. Kd tree k-means clustering uses a tree data structure with k dimensions in the process of initializing the cluster midpoint. Kd tree k means clustering is also a method capable of handling data sets that have noise / outliers by eliminating 20 percent of the cluster midpoint candidates with the lowest density value [11].

The cluster validity index is very important because it is designed with the aim of estimating how well a partition fits into the underlying dataset structure [13]. Separation is done by calculating the distance between the centers of the cluster used. However, this distance does not always reflect the quality of the partitions between clusters and sometimes gives confusing results [14]. In evaluating the clustering algorithm, two parameters are used to determine the clustering performance measure [15]. The first is a validation measure. Many cluster validation techniques are available [16]. In many indices, the validity of two cluster properties is taken into account, namely cohesiveness and separation [3]. The validity index cluster is used to find the best data partition. such as, for example Dunn [17], Davies-Bouldin (DB) in [18], [19] Davies-Bouldin index, Dunn index and Xie-Beni index, or Silhouette (SIL) [20].

This research will discuss about the application of the validity index on K-Means and Fuzzy C-Means in identifying the right number of clusters and the right partitioning technique. The grouping is done based on the similarity of its properties. A probabilistic (probability) approach can be useful in partition group analysis in order to obtain appropriate clustering strategies including data differences and random similarity relationships.

## 2. Theoritical basis

### A. KD tree

K-Dimensional Tree (KD-Tree) is a space partition data structure, and is a special case of a binary space partition tree. KD-Tree which aims to arrange the points in space with dimension k. In its implementation, KD-Tree is a binary tree where each node in the binary tree is a point with dimensions of k [21] In simple terms the properties of the k-d tree data structure can be described as follows [22]

a) The K-D tree is a binary tree.
b) The root node holds all elements of the data set.
c) The data elements at each node are divided into two parts based on certain variants, so that one part becomes the left subtree element and the other part becomes the right subtree element.
d) The left and right divisions will recursively divide again, as long as each node "h" does not satisfy
   a. $1 \leq h \leq w$

Referring to point (c) above, where the data set is divided into two parts using certain variants [22] used the main component as a variant.

According to Steven in [21], the KD-Tree K-Means Grouping algorithm workflow is as follows:

a) Build KD-Tree from a data set. The KD-Tree made has one bucket of leaves with a maximum of 20 data in the leaf bucket.
b) For each leaf bucket (L1, L2, ... Lj), calculate the density (Pj) for each leaf bucket Lj, and calculate the leaf bucket midpoint (Mj) by finding the average value of all the points on the leaf bucket Lj.
c) Choose the center of the first cluster C1 = Mz, where z = argmax (Pj).
d) For t = 2,...,K :
   a. For j = 1, ..., q, calculate the leaf bucket ranking value (Gj) with the formula: $G_j = \left\{ \begin{matrix} min \\ k = 0 \dots t \end{matrix} \left[ d\left(C_k . M_j\right) \right] \right\} \times P_j$ (1)
   b. Ct = Mz, where z = arg max (Gj)
e) Remove 20% of the bucket of leaves with the lowest density value. Return to step 3 and calculate the K position of the new cluster center point (c1, c2, ... ck).
f) Run the K-Means Clustering algorithm by initializing the midpoints (C1, ... Ck) and (c1, c2, ... ck).

### B. K-Means

The K-means algorithm is a partitioning algorithm, because K-Means is based on determining the initial number of groups by determining the initial centroid value [23]. The following is the flow of the K-Means algorithm [24]:

a) Determine the number of k, where k is the number of clusters to be formed. To determine the number of clusters k is done with several considerations such as theoretical and conceptual considerations that might be proposed to determine how many clusters. In this research, to determine k, the kd tree algorithm will be used

**JurnalMantik**is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

b) Determine the center point of the cluster randomly or randomly, the center point of the cluster is often referred to as the centroid.

c) After determining the initial centroid, each data will look for the nearest centroid, namely by calculating the distance of each data to each centroid using the correlation formula between two objects, that is *Euclidean Distance*.

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

Information:
$D_e$= Euclidean Distance
$i$ = the number of objects,
($x,y$)= are the coordinates of the object, and
($s,t$)= is the coordinates of the centroid (the center of the cluster)

d) Then allocate each object into the cluster based on the minimum distance. [25] Find the new centroid using the following equation

$$\bar{v}_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} x_{kj} \quad (3)$$

Information:
$\bar{v}$ = centroid / average of the ith cluster for the j-th variable.
Ni = centroid / average of the i-th cluster for the j-th variable.
$k$ = index of the cluster.
$j$ = Variable index.
$x_{kj}$ = the k-th data value in the cluster for the j-variable.

e) Go back to steps 3, 4, and 5. If in the second iteration there are no cluster members moving to another cluster, then the iteration stops, but if there are cluster members moving to another cluster then return to steps 3, 4 and 5 Do the next iteration until no cluster members move to another cluster [24]

## C. Fuzzy C-Means

Fuzzy C-Means discusses the concept of similarity in function of adjacent objects and finds the cluster center point as a prototype. For some data objects there is no limit to one class only, but the data can be grouped based on the degree of membership, namely between 0 and 1 which indicates partial data membership [26]. The following is the Fuzzy C-Means algorithm according to [27], that is:

a) Input data that will be grouped X, in the form of a matrix with size n x p (n = number of data samples, p = attribute of each data). Xkj = k-th sample data ($k$ = 1,2,…, $n$), jth attribute ($j$ = 1,2,3, .., $m$).
specify:
   a. Number of clusters = c
   b. Weight power=m
   c. Maximum iteration = *MaxIter*
   d. The smallest error expected = ξ
   e. initial objective function = P0 = 0
   f. first iteration = t = 1

b) Generates a random or random number (μik ,, i = 1,2, .., c; k = 1,2, .., n), as elements of the initial

$$\text{partition } \mathrm{U}U_0 = \begin{bmatrix} \mu_{11}(x_1) & \mu_{12}(x_2) & \cdots\mu_{1c}(x_c) \\ : & : & : \\ \mu_{11}(x_1) & \mu_{12}(x_2) & \ldots\mu_{1c}(x_c) \end{bmatrix} (4)$$

The partition matrix in fuzzy clustering must meet the following conditions:

$$\mu_{ik} = [0,1]; (1 \leq i \leq c; 1 \leq k \leq n)$$

$$\sum_{i=1}^{n} \mu_{ik} = 1;\ 1 \leq i \leq c$$

$$0 < \sum_{i=1}^{c} \mu_{ik<c}\ ;\ 1 \leq k \leq n \ (5)$$

Count the number of each column (attribute)

$$Q_j = \sum_{i=1}^{c}\left(\mu_{ik}\right) \ (6)$$

with j=1,2,3,…,m
then count :

$$\mu_{ik} = \frac{\mu_{ik}}{Q_j} \ (7)$$

c) Calculatethe cluster to K: Vij, where i = 1,2,3,…, c and j = 1,2,3,…, m

**JurnalMantik**is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

$$V_{ij} = \frac{\sum_{k=1}^{n}((\mu_{ik})^{m} * X_{kj})}{\sum_{k=1}^{n}(\mu_{i,k})m} \quad (8)$$

$$V = \begin{bmatrix} v_{11} & \cdots & v_{1m} \\ \vdots & \ddots & \vdots \\ v_{c1} & \cdots & v_{cm} \end{bmatrix} \quad (9)$$

d) Calculate the objective function in iteration t, Pt using the following equation:

$$P_t = \sum_{k=1}^{n} \sum_{i=1}^{c} \left( \left[ \sum_{j=1}^{m} (X_{kj} - V_{ij})^2 \right] (\mu_{ik})^m \right) \quad (10)$$

Calculate changes in the partition matrix:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{p}(X_{kj} - V_{ij})^2\right]^{\frac{-1}{p-1}}}{\sum_{i=1}^{c}\left[\sum_{j=1}^{p}(X_{kj} - V_{ij})^2\right]^{\frac{-1}{p-1}}} \quad (11)$$

e) Check the stop condition if

a. If (| Pt-pt-1 | <ξ) or (t <maximum iteration) then stop

b. If not then t = t + 1 then repeat step 4

### D. Davies Bouldin Index

David L. Davies and Donald W. Bouldin introduced a method named after them, the Davies-Bouldin Index (DBI) used to evaluate clusters [28]. Evaluation using the Davies Bouldin Index has an internal cluster evaluation scheme, where the results of the clusters are good or not seen from the quantity and closeness of cluster result data [29]. The stages of calculating the Davies Bouldin Index are as follows:

The stages of calculating the Davies Bouldin Index are as follows:

a) Sum of Square Within-cluster (SSW)

To determine the cohesion in the ith cluster is to calculate the value of the Sum of Square Within-cluster (SSW).

$$SSW_i = \frac{1}{m_i} \sum_{j=i}^{m_i} d(x_j, c_i) \quad (12)$$

b) *Sum of Square Between-cluster* (SSB)

The calculation of the Sum of Square Between Clusters (SSB) aims to determine the separation between clusters.

$$SSB_{i,j} = d(c_i, c_j) \quad (13)$$

c) *Ratio* (Rasio)

aims to determine the comparison value between cluster i and cluster j. To calculate the ratio value owned by each cluster, the following equation is used [28]

.
$$SSB_{i,j} = \frac{SSW_1 + SSW_J}{SSB_{i,j}} \quad (14)$$

d) *Davies Bouldin Index*

The ratio value obtained from the above equation (14) is used to find the Davies-Bouldin Index (DBI) value using the following equation:

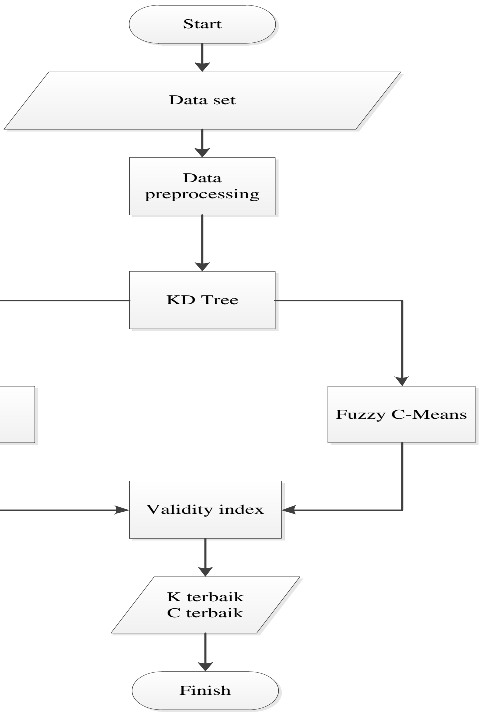
$$DB1 = \frac{1}{k} \sum_{i=1}^{k} max_{i=j}(R_{i,j}) \quad (15)$$

From this equation, k is the number of clusters. The smaller the DaviesBouldin Index (DBI) value obtained (non-negative> = 0), the better the cluster obtained from clustering using the clustering algorithm [29]

## 3. Metodologi

The first stage is the selection of the dataset to be processed, the dataset to be used in this study is a dataset of weekly sales transactions from https://archive.ics.uci.edu/ml/ which contains product sales for 52 weeks, then the data will be processed using the kd-tree method that will be used as the initial centroid for the clustering process uses the K-Means and Fuzzy C-Means algorithms. The next stage of processing results using the Kd tree method will be processed using two methods, namely: the K-Means clustering method and the Fuzzy C-Means clustering, then the resulting clusters will be validated using the Davies bouldin index method, so that it is expected to produce k and optimal c. The data processing process can be seen from the flow chart below:

**JurnalMantik**is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).



**Gambar 2.1** *Flowchart* kerangka kerja yang diusulkan

## 4. Result and discussion

Clustering is done by determining the initial centroid using the KD-Tree algorithm. Each class was tested using the K-Means and Fuzzy C-Means methods. The clustering process that has been carried out will then be validated using the Davies Bouldin index, where according to the provisions the smallest validation value is the best clustering result from other classes, the following clustering results can be seen in the following table:

**TABLE 1**

The results of clustering with leaf 5 and iteration max 100

| LEAF=5 IMAX=100 | K MEANS | | | | | | | | | | Fuzzy C-Means | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DBI | 0,45 | 0,63 | 0,79 | 1,48 | 1,46 | 1,6 | 1,96 | 2,31 | 2,52 | 2,75 | 0,5 | 0,63 | 1,06 | 1,7 | 1,22 | 2,5 | 1,4 | 1,49 | 2,36 | 2,23 |
| Time | 0,48 | 0,02 | 0,01 | 5,43 | 0,21 | 0,03 | 0,02 | 0,02 | 0,25 | 0,02 | 11,6 | 17 | 22,8 | 28,7 | 33,8 | 38,8 | 44,6 | 50,1 | 64 | 61,8 |
| Number of Iterasions | 7 | 7 | 7 | 9 | 24 | 12 | 14 | 12 | 14 | 12 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

**TABLE 2**

The results of clustering with leaf 10 and iteration max 100

| LEAF=10, I MAX=100 | K MEANS | | | | | | | | | | Fuzzy C-Means | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DBI | 0,45 | 0,63 | 0,79 | 1,48 | 1,64 | 2,02 | 2,26 | 2,41 | 2,7 | 2,84 | 0,5 | 0,63 | 1,06 | 1,7 | 2,75 | 2,67 | 4,71 | 3,49 | 3,17 | 3,07 |
| Time | 0,25 | 0,01 | 0,01 | 0,03 | 0,01 | 0,02 | 0,02 | 0,03 | 0,03 | 0,03 | 12,1 | 17 | 22,7 | 28,3 | 33,5 | 39,1 | 44,4 | 51 | 55,5 | 61,3 |
| Number of iterations | 7 | 7 | 6 | 21 | 12 | 11 | 13 | 13 | 13 | 20 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

JurnalMantikis licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

**TABLE 3**
The results of clustering with leaf 15 and iteration max 100

| LEAF=15, I MAX=100 | K MEANS | | | | | | | | | | Fuzzy C-Means | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DBI | 0,45 | 0,63 | 0,86 | 0,92 | 1,46 | 1,6 | 1,7 | 1,96 | 2,26 | 2,41 | 0,5 | 0,63 | 1,06 | 1,47 | 1,2 | 3,32 | 3,23 | 2,42 | 1,73 | 2,23 |
| time | 0,01 | 0,02 | 0,01 | 0,03 | 0,04 | 0,02 | 0,03 | 0,03 | 0,02 | 0,04 | 11,6 | 17,1 | 23,1 | 28,2 | 33,4 | 39,2 | 44,9 | 51,2 | 56,2 | 61,7 |
| Number of iterations | 7 | 6 | 7 | 6 | 24 | 15 | 14 | 19 | 14 | 24 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

**TABLE 4**
The results of clustering with leaf 20 and iteration max 100

| LEAF=20, I MAX=100 | K MEANS | | | | | | | | | | Fuzzy C-Means | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DBI | 0,45 | 0,63 | 0,86 | 0,92 | 1,46 | 1,6 | 1,7 | 1,96 | 2,26 | 2,41 | 0,5 | 0,63 | 1,06 | 1,47 | 1,2 | 3,32 | 3,23 | 2,42 | 1,73 | 2,23 |
| Time | 0,01 | 0,02 | 0,01 | 0,01 | 0,03 | 0,03 | 0,02 | 0,03 | 0,03 | 0,26 | 11,6 | 17,2 | 22,8 | 27,8 | 33,5 | 39,2 | 45,2 | 50,3 | 56,1 | 61,8 |
| number of iterations | 7 | 8 | 7 | 6 | 24 | 15 | 14 | 19 | 14 | 24 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

The following are the results of clustering with max 200 iterations carried out on data with different number of leaf which can be seen in the following table:

**TABLE 5**
The results of clustering with leaf 5 and iteration max 200

| LEAF=5, I MAX=200 | K MEANS | | | | | | | | | | Fuzzy C-Means | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DBI | 0,45 | 0,63 | 0,79 | 1,48 | 1,46 | 1,6 | 1,96 | 2,31 | 2,52 | 2,75 | 0,5 | 0,63 | 1,06 | 1,7 | 1,22 | 2,5 | 1,2 | 3,49 | 2,36 | 3,98 |
| Time | 0,02 | 0,20 | 0,02 | 0,02 | 0,03 | 0,31 | 0,2 | 0,02 | 0,02 | 0,03 | 24,6 | 35,9 | 47,3 | 60,5 | 67 | 77,9 | 94,7 | 101 | 111 | 123 |
| number of iterations | 7 | 7 | 6 | 9 | 24 | 12 | 14 | 12 | 14 | 12 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 |

**TABLE 6**
The results of clustering with leaf 10  and iteration max 200

| LEAF=10, I MAX=200 | K MEANS | | | | | | | | | | Fuzzy C-Means | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DBI | 0,45 | 0,63 | 0,79 | 1,48 | 1,64 | 2,02 | 2,26 | 2,41 | 2,7 | 2,84 | 0,5 | 0,63 | 1,06 | 1,7 | 2,75 | 2,67 | 4,76 | 2,42 | 3,17 | 3,07 |
| Time | 0,01 | 0,15 | 0,02 | 0,03 | 0,02 | 0,02 | 0,31 | 2,64 | 0,03 | 0,03 | 24,1 | 36 | 48,4 | 56,5 | 66,7 | 77,9 | 94,9 | 156 | 111 | 123 |
| number of iterations | 7 | 7 | 6 | 21 | 12 | 11 | 13 | 13 | 13 | 20 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 |

**TABLE 7**
The results of clustering with leaf 15 and iteration max 200

| LEAF=15, I MAX=200 | K MEANS | | | | | | | | | | Fuzzy C-Means | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DBI | 0,45 | 0,63 | 0,86 | 0,92 | 1,46 | 1,6 | 1,7 | 1,96 | 2,26 | 2,4 | 0,5 | 0,63 | 1,06 | 1,46 | 1,2 | 3,18 | 3,23 | 2,42 | 1,73 | 2,23 |
| Time | 0,01 | 0,06 | 0,02 | 0,04 | 0,31 | 0,03 | 0,02 | 0,03 | 0,02 | 0,03 | 24,2 | 35,8 | 48,1 | 56 | 66,9 | 78,5 | 95,5 | 101 | 112 | 123 |
| number of iterations | 7 | 8 | 7 | 6 | 24 | 15 | 14 | 19 | 14 | 24 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 |

**TABLE 8**
The results of clustering with leaf 20 and iteration max 200

| LEAF=20, I MAX=200 | K MEANS | | | | | | | | | | Fuzzy C-Means | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DBI | 0,45 | 0,63 | 0,86 | 0,92 | 1,46 | 1,6 | 1,7 | 1,96 | 2,26 | 2,41 | 0,5 | 0,63 | 1,06 | 1,46 | 1,2 | 3,18 | 3,23 | 2,42 | 1,73 | 2,23 |
| Time | 0,02 | 0.15 | 0,02 | 0,02 | 0,02 | 0,03 | 0,02 | 0,06 | 0,03 | 0,24 | 24,3 | 35,9 | 48,2 | 26,2 | 67,5 | 78,3 | 95,2 | 101 | 112 | 113 |
| number of iterations | 7 | 8 | 7 | 6 | 24 | 15 | 14 | 19 | 14 | 24 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 |

**JurnalMantik**is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

The following are the results of clustering with max 300 iterations carried out on data with different number of leaf which can be seen in the following table:

**TABLE 9**
The results of clustering with leaf 5 and iteration max 300

| LEAF=5, I MAX=300 | **K MEANS** | | | | | | | | | | **Fuzzy C-Means** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DBI | 0,45 | 0,63 | 0,79 | 1,48 | 1,6 | 1,6 | 1,96 | 2,31 | 2,52 | 2,75 | 0,5 | 0,63 | 1,06 | 1,7 | 1,22 | 2,5 | 1,2 | 1,49 | 2,36 | 3,98 |
| Time | 0,02 | 0,04 | 0,02 | 0,02 | 0,03 | 0,02 | 0,02 | 0,03 | 0,02 | 0,02 | 35,1 | 55 | 74,3 | 93,1 | 104 | 121 | 139 | 156 | 174 | 191 |
| number of iterations | 7 | 7 | 6 | 9 | 24 | 12 | 14 | 12 | 14 | 12 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 |

**TABLE 10**
The results of clustering with leaf 10 and iteration max 300

| LEAF=10, I MAX=300 | **K MEANS** | | | | | | | | | | **Fuzzy C-Means** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DBI | 0,45 | 0,63 | 0,79 | 1,48 | 1,64 | 2,02 | 2,26 | 2,41 | 2,7 | 2,84 | 0,5 | 0,63 | 1,06 | 1,7 | 2,75 | 2,67 | 4,76 | 3,49 | 3,17 | 3,07 |
| Time | 0,02 | 0,23 | 0,02 | 0,03 | 0,02 | 0,02 | 0,02 | 0,03 | 0,02 | 0,03 | 35,2 | 50,7 | 67,1 | 84 | 99,5 | 117 | 133 | 158 | 173 | 190 |
| number of iterations | 7 | 7 | 6 | 21 | 12 | 11 | 13 | 13 | 13 | 20 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 |

**TABLE 11**
The results of clustering with leaf 15 and iteration max 300

| LEAF=15, I MAX=200 | **K MEANS** | | | | | | | | | | **Fuzzy C-Means** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DBI | 0,45 | 0,63 | 0,86 | 0,92 | 1,46 | 1,6 | 1,7 | 1,96 | 2,26 | 2,41 | 0,5 | 0,63 | 1,06 | 1,46 | 1,2 | 3,18 | 3,23 | 2,42 | 1,73 | 2,23 |
| Time | 0,02 | 0,06 | 0,02 | 0,02 | 0,03 | 0,02 | 0,03 | 0,03 | 0,02 | 0,03 | 35,2 | 50 | 74,5 | 90,4 | 104 | 122 | 140 | 156 | f173,56 | 191 |
| number of iterations | 7 | 8 | 7 | 6 | 24 | 15 | 14 | 19 | 14 | 24 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 |

**TABLE 12**
The results of clustering with leaf 5 and iteration max 300

| LEAF=20, I MAX=200 | **K MEANS** | | | | | | | | | | **Fuzzy C-Means** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DBI | 0,45 | 0,63 | 0,86 | 0,92 | 1,46 | 1,6 | 1,7 | 1,96 | 2,26 | 2,41 | 0,5 | 0,63 | 1,06 | 1,46 | 1,2 | 3,18 | 3,23 | 2,42 | 1,73 | 2,23 |
| Time | 0,02 | 0,02 | 0,02 | 0,02 | 0,03 | 0,03 | 0,03 | 0,02 | 0,02 | 0,03 | 34,5 | 50,7 | 67,1 | 83,7 | 101 | 117 | 144 | 156 | 174 | 191 |
| number of iterations | 7 | 8 | 7 | 6 | 24 | 15 | 14 | 19 | 14 | 24 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 | 300 |

Based on the test results that we can see in the table above, optimal k is at k = 2 and optimal c is at c = 2, the full results can be seen in the table below:

**TABLE 13**
The best number of cluster members with different leaves and iterations

| **iterasi** | **5** | | | | **10** | | | | **15** | | | | **20** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **k means** | | **fuzzy c means** | | **k means** | | **fuzzy c means** | | **k means** | | **fuzzy c means** | | **k means** | | **fuzzy c means** | |
| 100 | 665 | 146 | 647 | 164 | 665 | 146 | 647 | 164 | 665 | 146 | 647 | 164 | 665 | 146 | 647 | 164 |
| 200 | 665 | 146 | 647 | 164 | 665 | 146 | 647 | 164 | 665 | 146 | 647 | 164 | 665 | 146 | 647 | 164 |
| 300 | 665 | 146 | 647 | 164 | 665 | 146 | 647 | 164 | 665 | 146 | 647 | 164 | 665 | 146 | 647 | 164 |

The results of the analysis based on the table above can be seen that the overall number of members of the cluster 2 with the number of different leaves and the maximum iteration produces the same number of members for the different number of leaves, while for k means the index method 0 = 665 and index 1 = 146 while for the fuzzy c method means index 0 = 647 and index 1 = 164. Overall, the selection of the initial

**JurnalMantik**is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

centroid using the kd tree makes the initial cluster center consistent so that the data grouping is stable, both the results of clustering using the k means algorithm and the Fuzzy C means algorithm.

The analysis obtained is based on the test results above, that is:

a) The clustering results obtained tend to have almost the same dbi value, both from the results of K-Means and Fuzzy C-Means clustering.
b) The amount of data on the leaf in determining the initial centroid using the KD tree algorithm affects the validation results and the number of iterations.
c) The results of the tests conducted show that the maximum iteration also affects the validation results and the clustering process time for the Fuzzy C-Means method.

## 5. Conclusion

a) Based on the test results using the K means algorithm and Fuzzy C Means algorithm, with the determination of the initial centroid using the kd tree, similar results were obtained between the two algorithms, where the optimal K and C were obtained for sales transaction data, a weekly dataset from 2 clusters, with the Davies Bouldin index. (Dbi) K means = 0.45 while Dbi Fuzzy C means = 0.5.
b) The results of clustering tend to be the same, where the best number of cluster members is cluster 2(two).With different number of leaves and different maximum iterations produce the same number of members, where for the k method means index 0 = 665 members and index 1 = 146 members while for the fuzzy method c means index 0 = 647 members and index 1 = 164. Overall, the selection of the initial centroid using the kd tree algorithm makes the initial cluster center consistent so that the data grouping is stable, both the results of clustering using the k means algorithm and the fuzzy c means algorithm even though the max iteration is different.

## 6. References

[1] Xu & Wunsch, 2009, Clustering, pp. 1-15, Wiley-IEEE Press, Available from: Ebook Library. [Agustus 2009].

[2] Hämäläinen J, Jauhiainen S & Kärkkäinen T 2017, Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering. Algorithms 2017, 10, 105.

[3] Halkidi et., al 2001, On Clustering Validation Techniques, Journal of Intelligent Information Systems, 17:2/3, pp. 107–145, 2001.

[4] Han et., al 2012, Data Mining Concepts and Techniques. 3rd Edition, Morgan Kaufmann Publishers, Waltham.

[5] Kusumadewi S, Hartati S, Harjoko, Agus, Wardoyo & Retantyo 2006. Fuzzy Multi-Attribute Decision Making (FUZZY MADM). Graha Ilmu, Yogyakata.

[6] Min & Kai-fei 2015, Improved research to k-means initial cluster centers, 978-1-4673-9295-2/15 $31.00 © 2015 IEEE, 2015 Ninth International Conference on Frontier of Computer Science and Technology.

[7] Ding et al., 2015, Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm, Neurocomputing, 2015, 188:233-238.

[8] Arbelaitz O, Gurrutxaga I, Muguerza J, Perez J.M & Perona I. “An extensive comparative study of cluster validity indices”, Pattern Recognition 46. 2013. Pp 243-256.

[9] Zahra S, Ghazanfar M.A, Khalid A, Azam, M.A Naeem, N & Prugel-Bennett A. “Novel centroid selection approaches for KMeans-clustering based recommender systems”, Information Sciences, 320 .2015. pp 156–189.

[10] Nazeer K.A & Sebastian M.P 2010, Clustering biological data using enhanced k-means algorithm, in: Electronic Engineering and Computing Technology, Springer, 2010, pp. 433–442 (chapter 37).

[11] Stephen J Redmond and Conor Heneghan, "A method for initialising the K-means clustering algorithm using kd-trees," Pattern Recognition Letters, vol. 28, no. 8, pp. 965–973, June 2007.

[12] aI. Katsavounidis, C.C.J. Kuo, and Z. Zhen, "A new initialization technique for generalized lloyd iteration," IEEE Signal Processing Letter, vol. 1, no. 10, pp. 144–146, 1994.

[13] Lianyu H & Caiming Z 2019, An Internal Validity Index Based On density-involved distance ,IEEE Access, VOLUME 4, 2016, no. 1, pp. 1-14, 2019.

[14] Said, R. Hadjidj & S. Foufou, Cluster validity index based on jeffrey divergence, Pattern Analysis and Applications, vol. 20, no. 1, pp. 21–31, 2017.

[15] Salem et al., A vertex chain code approach for image recognition. ICGST International Journal on Graphics, vision and Image processing 05 (2005).

[16] Zhao et., al 2012, Imagination Difficulty and New Product Evaluation, J PROD INNOV MANAG 2012;29 (S1):pp. 76–90.

[17] Dunn JC 1974, Well separated clusters and optimal fuzzy partitions.J Cybern 4:95–104.

[18] Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1(4):224–227.

[19] Pakhira MK, Bandyopadhyay S, Maulik U (2004) Validity index for crisp and fuzzy clusters. Pattern Recogn 37(3):487–501.

**JurnalMantik**is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

[20] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” Journal of computational and applied math- ematics, vol. 20, pp. 53–65, 1987.

[21] Soelaiman Isye; Gosno, Eric Budiman, R. A. (2013) ‘Implementasi KD-Tree K-Means Clustering untuk Klasterisasi Dokumen’, Jurnal Teknik ITS, 2(Vol 2, No 2 (2013)), pp. A432–A437. Available at: http://ejurnal.its.ac.id/index.php/teknik/article/view/3872.

[22] Likas, A., Vlassis, N., Verbeek, J.J., 2003. The global k-means clustering algorithm. Pattern Recognition 36, 451–461.

[23] Madhulatha, S. T 2012, “An overview on clustering method”, IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-725.

[24] Poteras, C.M., Mihăescu, M.C. & Mocanu, M. 2014. An optimized version of the kmeans clustering algorithm. Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, pp. 695–699.

[25] Goyal, M. & Kumar, S. 2014. Improving the initial centroids of k-means clustering algorithm to generalize its applicability. Journal of The Institution of Engineers 95(4): 345–350.

[26] Mr.Kaushi K Phukon MCA, P. H. K. B. (2013) ‘Extension of the Fuzzy C Means Clustering Algorithm To Fit With the Composite Graph’, International Journal Of Cognitive Research In science,engineering and education, 1(2).

[27] Andriyani, T. M., Linawati, L., and Setiawan, A., 2013, "Penerapan Algoritma Fuzzy C-Means (Fcm) Pada Penentuan Lokasi Pendirian Loket Pembayaran Air PDAM Salatiga," Prosiding Seminar Nasional Sains dan Pendidikan Sains VIII, Fakultas Sains dan Matematika Universitas Kristen Satya Wacana, Salatiga.

[28] Nawrin, S., Rahman, M.R. & Akhter, S. 2017. Exploreing k-means with internal validity indexes for data clustering in traffic management system. International Journal of Advanced Computer Science and Applications 8(3): 264-272.

[29] Bates A. & Kalita J. 2016. Counting clusters in twitter posts. Proceedings of the 2nd International Conference on Information Technology for Competitive Strategies, pp. 85.

**JurnalMantik**is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).