Application of Fuzzy C-Means Algorithm for Clustering Customers

Oktaviana Nirmala Purba \(^1\), Dian Novianti Sitompul \(^2\), Tua Holomoan Harahap \(^3\), Sri Rahmah Dewi Saragih \(^1\), Rizka Fahruza Siregar \(^4\)

\(^1\) Department of Mathematics Education, Faculty of Teacher Training and Education, Universitas Asahan, Kisaran, 21216, North Sumatra, Indonesia
\(^2\) Department of Accounting Education, Faculty of Teacher Training and Education, Universitas Muhammadiyah Sumatera Utara, Medan, 20238, North Sumatra, Indonesia
\(^3\) Department of Mathematics Education, Faculty of Teacher Training and Education, Universitas Muhammadiyah Sumatera Utara, Medan, 20238, North Sumatra, Indonesia
\(^4\) Department of Civil Engineering, Faculty of Engineering, Universitas Pembinaan Masyarakat Indonesia, Medan, 20227, North Sumatera, Indonesia

ABSTRACT

Today's fierce business competition requires companies to focus on the needs desired by consumers. This makes companies have to think about how to manage customer data so that it can be utilized properly for the development of marketing strategies. And grouping (cluster) customers based on their respective characteristics can be an alternative in solving these problems. In clustering (cluster) customers, there are several data mining clustering methods that can be used, one of which is the Fuzzy C-Means (FCM) method. FCM is a clustering algorithm where one object can be a member of several clusters and FCM cluster boundaries are vague. The output of FCM is a row of cluster centers and several degrees of membership for each data point. In this clustering, customers will be divided into 4 customer clusters namely Golden, Silver, Bronze, and Iron with the variables used as a reference are the final purchase date, purchase frequency and total purchase. The data used is customer transaction data for the period September - December 2015. The total data is 709 transactions from 75 customers. After the data is processed with the Fuzzy C-Means method, the final results show that the iteration ends at the 30th iteration with a change in the objective function of 9.8. The resulting customer clusters are Golden: 27, Silver: 15, and Bronze: 33 with a cluster validity of 0.596277.

KEYWORDS
Data Mining; Fuzzy C-Means (FCM); Customer Cluster; Cluster Validity; Transaction, RFM

INTRODUCTION

Today's intense business competition requires companies to focus on the needs desired by consumers [1]. Customers are important assets in the company that must be well maintained, especially potential customers. With the large number of customers in a company needs a strategy to determine the company's potential customers by grouping customers [2].

Grouping customers based on their respective characteristics will affect the marketing management of a company. Therefore, to group customers based on their respective characteristics, one of them needs a data mining method. Data mining is one of the sciences in the field of informatics that studies data mining and text documents are one of the mined sciences. Clustering is one of the techniques of one of the data mining functionalities, clustering algorithm is an algorithm for grouping a number of data into certain data groups (clusters) [4]. There are many algorithms used in clustering, one of which is Fuzzy C-Means (FCM). FCM is a clustering algorithm where one object can be a member of several clusters and FCM cluster boundaries are vague. The basic concept of FCM is to first determine the center of the cluster. And each data point has a membership degree for each cluster. The value of the membership degree in the FCM algorithm is between 0 and 1 [5].
In this study, researchers will take a research case study at CV. Mataram Jaya Bawen. CV. Mataram Jaya Bawen is one of the dealers and workshops that provides service services and product sales for motorized vehicles. Product sales at Mataram Jaya Bawen have been recorded in the product sales database to be used as a daily report on sales results. But the sales reports that are owned have not been maximally utilized by the company. One form of utilization of sales data is by making customer clustering to find out potential customers for the company. Potential customers must be maintained because potential customers provide great benefits for the company. Therefore, the creation of customer clustering is needed by the company to determine specific marketing strategies to retain CV. Mataram Jaya customers.

To determine the cluster that suits CV Mataram Jaya customers, customers are assessed for their profitability to the company from transactions that have been made using the RFM (Recency, Frequency, and Monetary) method. Recency is the length of the time interval since the customer made the last transaction, frequency is related to the level of frequency of customers making transactions, and Monetary is the value of transactions made during a certain period. Based on the explanation of the problems found at CV Mataram Jaya Bawen, the researcher chose the title “Application of the Fuzzy C-Means (FCM) Algorithm for customer clustering at CV. Mataram Jaya Bawen”.

**METHOD**

In this research, the research method used by researchers is the standard data mining process model CRISP-DM (Cross Industry Standard process) developed in 1996 by analysts from several industries providing data mining processes as a general problem-solving strategy for research. In CRISP-DM, a data mining project has the following steps:

**Business Understanding**
CV. Mataram Jaya Bawen is a company that sells motorcycle spare parts products. This company has dozens of types of products offered. The products are sold wholesale and retail. The current condition of CV. Mataram Jaya Bawen has more than 75 regular customers and monthly sales turnover reaches tens of millions of rupiah.

**Data Understanding**
In this phase, it is necessary to understand the needs of data related to achieving data mining objectives. From the data retrieval process, 700 sales data of spare parts products for the period September - December 2015 were obtained, which were purchased by customers wholesale and retail. The data includes note number, sales date, item name, quantity, and net total. And below is an example of the original data:

| KodePangkat | NomorNota | Tgl | Banyak | NamaBeliing | Type | Mark | Diskon | HargaSatuan | HargaSembol | TotalNota | TotalDiskon | TotalNetto |
|-------------|------------|-----|--------|-------------|------|------|-------|-------------|-------------|-----------|------------|------------|------------|
| YH0001      | F820220004900 | 06/10/2023 | 1 PC | KLV I-41    | 74 x 1300 | FEDEAL | 0.00   | 33000       | 33000       | 0         | 33000       | 33000      |
| YH0002      | F820122004900 | 06/10/2023 | 2 PC | TOLIAT 25/165 | 1 x 3900 | YAMALUBE | 28000 | 28000       | 0         | 28000     | 0           | 28000      |
| YH0003      | F820122005200 | 06/10/2023 | 1 PC | GTI Ø80 | V iXON | GENIEN | 50000 | 50000       | 0         | 50000     | 0           | 50000      |
| YH0004      | F820122005200 | 06/10/2023 | 1 PC | YAMA MOTOR BIKE | V iXON | GENIEN | 50000 | 50000       | 0         | 50000     | 0           | 50000      |
| YH0005      | F820122005200 | 06/10/2023 | 1 PC | YAMA MOTOR BIKE | V iXON | GENIEN | 50000 | 50000       | 0         | 50000     | 0           | 50000      |
| YH0006      | F820122005200 | 06/10/2023 | 1 PC | YAMA MOTOR BIKE | V iXON | GENIEN | 50000 | 50000       | 0         | 50000     | 0           | 50000      |
| YH0007      | F820122005200 | 06/10/2023 | 1 PC | YAMA MOTOR BIKE | V iXON | GENIEN | 50000 | 50000       | 0         | 50000     | 0           | 50000      |
| YH0008      | F820122005200 | 06/10/2023 | 1 PC | YAMA MOTOR BIKE | V iXON | GENIEN | 50000 | 50000       | 0         | 50000     | 0           | 50000      |
| YH0009      | F820122005200 | 06/10/2023 | 1 PC | YAMA MOTOR BIKE | V iXON | GENIEN | 50000 | 50000       | 0         | 50000     | 0           | 50000      |
| YH0010      | F820122005200 | 06/10/2023 | 1 PC | YAMA MOTOR BIKE | V iXON | GENIEN | 50000 | 50000       | 0         | 50000     | 0           | 50000      |

Figure 1. Original Data

**Data Preparation**
In this phase, the data will be prepared so that it is easy to mine. The preparation process will include 3 basic things, namely:

- **To be continued...**

https://doi.org/10.56211/hanif.v1i1.8
1. Data Selection
Data selection is the selection of data used in the data mining process. In this process, there is also a selection of attributes that are tailored to the data mining process.

2. Data Preprocessing
Data preprocessing is the quality assurance of data that has been selected previously in data selection. In this phase, the problems that exist are noisy data and missing values.

3. Data Transformation
Data Transformation is a grouping of attributes or fields that have been selected into 1 table by denormalizing.

Below are the results of the selection of attributes needed in the customer clustering process using the RFM method, namely attributes related to recency, frequency, and monetary.

Table 1. Selection of CV.Mataram Jaya Bawen Attributes

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Code</td>
<td>Indicates recency, which is the date of the last purchase transaction made by the customer. Customer.</td>
</tr>
<tr>
<td>Purchase Distance</td>
<td>Marking frequency, which is the number of transactions during the period. specified</td>
</tr>
<tr>
<td>FrequencyBuy</td>
<td>Signaling monetary, is the amount of money during the period specified</td>
</tr>
</tbody>
</table>

**Modeling**
From the previous data, the dataset is processed by weighting according to the value domain to facilitate data processing before being modeled with the Fuzzy C-Means algorithm so as to produce membership degree data for each variable for determining consumer labels, namely customer classes.

Table 2. Value Domain

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Value Domain</th>
<th>Category Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recency</td>
<td>R ≥ 22 days</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>21 ≤ r ≤ 15 days</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>14 ≤ r ≤ 8 days</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>0 ≤ r ≤ 7 days</td>
<td>4</td>
</tr>
<tr>
<td>Frequency</td>
<td>0 ≤ f ≤ 4 transactions</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5 ≤ f ≤ 8 transactions</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>9 ≤ f ≤ 15 transactions</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>f ≥ 16 transactions</td>
<td>4</td>
</tr>
<tr>
<td>Monetary</td>
<td>0 ≤ m ≤ 5 millionrupiah</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5.1 ≤ m ≤ 10 millionrupiah</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>10.1 million ≤ f ≤ 15 million transactions</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>m ≥ 15.1 millionrupiah</td>
<td>4</td>
</tr>
</tbody>
</table>
From the data set above, data processing will be carried out using the FCM algorithm in accordance with the attributes used in clustering CV Mataram Jaya Bawen customers. The following is the application of attributes in the product sales dataset to determine the type of customer. Customer types are divided into 4, namely: Golden, Silver, Bronze, and Iron:

<table>
<thead>
<tr>
<th>Customer Code</th>
<th>R</th>
<th>F</th>
<th>M</th>
<th>Customer Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>YMH001</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>GOLDEN</td>
</tr>
<tr>
<td>YMH002</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>GOLDEN</td>
</tr>
<tr>
<td>YMH003</td>
<td>3</td>
<td></td>
<td>2</td>
<td>GOLDEN</td>
</tr>
<tr>
<td>YMH004</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>IRON</td>
</tr>
<tr>
<td>YMH005</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>IRON</td>
</tr>
<tr>
<td>YMH006</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>BRONZE</td>
</tr>
<tr>
<td>YMH007</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>SILVER</td>
</tr>
<tr>
<td>YMH008</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>SILVER</td>
</tr>
<tr>
<td>YMH009</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>IRON</td>
</tr>
<tr>
<td>YMH010</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>IRON</td>
</tr>
</tbody>
</table>

**Evaluation**
Evaluation of the model used by mining the dataset in a certain period. The clustering process is tested with various parameter values of the clustering algorithm. If it is not able to read the cluster correctly, it will return to the modeling process to improve the structure of the model used. The evaluation phase is considered complete if the business understanding has been answered properly.

**Deployment**
In this phase, the data mining application will be made. In this research, the deployment process is to perform the data mining process on the entire product sales database and no integration with the existing system at CV. Mataram Jaya Bawen.

Previous research used as a reference by the author to support this research includes clustering student grade data to group major concentrations using fuzzy c-means [5]. Here the fuzzy c-means algorithm is used to group the concentration of student majors according to the acquisition of their academic grades. This is useful for aligning students according to their expertise so that there is no mismatch with the concentration of the currently selected major. Data analysis carried out in this study uses the help of Matlab applications to form data clusters that are as expected. The results of this study are in the form of three cluster data that can be used for decision support for determining the concentration of 126 student data.

The next research is about grouping information systems students based on In the study, the algorithm used was fuzzy k-means to be able to see the level of academic ability of students according to the desired parameters and wanted to be seen by the caprodi and lecturers. Course weighting with the fuzzy k-means method in this study is largely determined by the value of the course attributes and the predicted value of each course. The process of iteration or repetition in the fuzzy k-means method is very important because the weight of the course is still very likely to change, therefore it continues to be repeated until it produces a fixed value so that it can affect the level of student ability because the level of student ability is seen based on the value of the course multiplied by the weight of each course so that the results of the grouping are determined by the weight of the course. And the result of this research is the grouping of students based on certain criteria such as gender, class and region of origin for competency levels above average, average, or below average.

Next research conducted by Megawati, Mukid, and Rahmawati on the use of the fuzzy c-means algorithm for market segmentation. [Here the researchers used the fuzzy c-means algorithm to group consumers into 2 clusters based on 10...
psychographic variables. The research data was obtained through distributing research questionnaires at RITA pasaraya Cilacap. The results of this study show consumer segmentation which is divided into 2 clusters. Respondents in cluster 1 pay attention to the level of low prices, completeness of goods, large discounts, satisfactory shopping services, strategic location, spacious parking, comfort when shopping, adequate public facilities, complete payment facilities, and cleanliness of the room compared to respondents in cluster 2. With this research, the right target market can be applied to a supermarket according to the case study taken by the researcher.

Based on the research references above, the researcher wants to develop the application of the fuzzy c-means (fcm) algorithm at CV. Mataram Jaya Bawen for clustering customers according to their respective criteria to facilitate the determination of the right marketing strategy. Later the data that will be processed by researchers is purchase data made by customers during a certain period to determine variables.

**Fuzzy C-Means (FCM)**

Clustering with the Fuzzy C-Means (FCM) method is based on fuzzy logic theory. This theory was first introduced by Lothfi Zadeh (1965) with the name fuzzy set. In fuzzy theory, the membership of a data is not given a value strictly with a value of 1 (being a member) and 0 (not looking for a member), but with a value of a membership degree value whose value range is 0 to 1.

The sum of the membership degree values of each data xi is always 1, which is formulated in the following equation:

\[ \sum_{j=1}^{k} u_{ij} = 1 \]  

(1)

Cluster cj can be formulated as follows:

\[ 0 < \sum_{j=1}^{k} u_{ij} < n \]  

(2)

The value of the membership degree of data xi in cluster cj, is formulated in the following equation:

\[ u_{ij} = \frac{D(x_i, c_j)^{w-2}}{\sum_{i=1}^{k} D(x_i, c_j)^{w-2}} \]  

(3)

Parameter \( c_j \) is the centroid of the jth cluster, \( D(x_i, c_j) \) is the distance between the data and the centroid, while \( w \) is the weighting exponent introduced in FCM. \( w \) has no fixed value, usually the value of \( w > 1 \) and is generally given a value of 2.

The membership value is stored in an \( N \times k \) fuzzy pseudo-partition matrix, where the rows are data, while the columns are membership values for each cluster:

\[ U = \begin{bmatrix} u_{11} [x_1] & u_{12} [x_1] & \cdots & u_{1k} [x_1] \\ u_{21} [x_2] & u_{22} [x_2] & \cdots & u_{2k} [x_2] \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} [x_n] & u_{n2} [x_n] & \cdots & u_{nk} [x_n] \end{bmatrix} \]  

(4)

To calculate the centroid in cluster1 for feature j, the following equation is used:

\[ c_{ij} = \frac{\sum_{i=1}^{N} (u_{ij})^w x_{ij}}{\sum_{i=1}^{N} (u_{ij})^w} \]  

(5)
The parameter N is the number of data, w is the weight of the rank, and is the membership degree value of data xi to cluster c1. While the objective function uses the following equation:

\[
J = \sum_{i=1}^{N} \sum_{j=1}^{k} (\mu_{ij})^w \| x_i - c_j \|^2
\]

(6)

Fuzzy C-Means Clustering Algorithm:
1. Initialization: determine the number of clusters (k ≥ 2), determine the weight rank (w > 1), determine the maximum number of iterations, determine the threshold for changing the objective function value (if necessary also change the centroid value).
2. Give the initial value to the fuzzy pseudo-partition matrix.
3. Perform steps 4 through 5 as long as the following conditions are met: (1) if the change in the objective function value is still above the specified threshold value; or (2) the change in the centroid value is still above the specified threshold value; or (3) the maximum iteration has not been reached.
4. Calculate the centroid value of each cluster. Recalculate the fuzzy pseudo partition matrix (the degree of membership of each data in each pseudo partition).

RESULTS AND DISCUSSION

The implementation of this algorithm contains information from the system that has been created and implemented in a web-based system. And here is a screenshot of the program that has been built by researchers:

Figure 2. Customer Data Input Page

On this page, the admin can add new customer or customer data owned by CV. Mataram Jaya Bawen. The customer data inputted is information about the name, contact and address of the customer.

Figure 3. DataCustomer page

On this page, the admin can view customer or customer data owned by CV. Mataram Jaya Bawen. Here the admin can also edit customer information through the edit menu.

https://doi.org/10.56211/hanif.v1i1.8
On this page the admin can input transaction data in the form of customer id pelanggan, purchase date, and total purchases made by customers during a certain period.

On this page, the admin can view transaction data that has been inputted previously. This transaction data will later be processed by the system for the calculation of the Fuzzy C-Means (FCM) algorithm.

On this page, the admin can add customer transaction data from an excel file, making it easier for the admin to process customer transaction data without having to input the transactions one by one.
On this page, the weighting process of customer transaction data will be carried out and data updates will be made to the databases before the calculation process is carried out by the system. If the weighting and calculation process is successful, a notification will appear as shown in Figure 4.7.
On this page, the system will display the results of the Fuzzy C-Means (FCM) algorithm calculation from the customer transaction data that has been processed. The system will display the algorithm calculation from iteration one to the last iteration.

On this page, the system will display a list of customers along with the cluster that matches the customer. The customer list includes customer id_pelanggan, customer name, maximum objective value and the corresponding cluster.

This page displays a graph showing the number of customers in a cluster, making it easier for companies to monitor customers. This research uses the Fuzzy C-Means (FCM) method in processing transaction data carried out by customers. the calculation of this algorithm consists of initializing data in the form of membership data, calculating membership data, and calculating membership data.

Centroid, calculate the distance of the data to the centroid, calculate the membership value, and calculate the objective function value. The objective function value affects the iterations performed, because if the change in the objective function value has not reached the smallest positive value, iterations will continue. After the objective function value
reaches the smallest positive value, then the data cluster can be determined. The final cluster formed shows that the number of golden = 28, silver 14, and bronze 33 customers.
To determine the validity of the cluster in the formation of customer clusters in the system, researchers calculate the partition coefficient (PC) as an evaluation of the data membership value in each cluster. The greater value (close to 1) means that the quality of the cluster obtained is getting better. the result of the PCI Index to calculate the accuracy of the cluster is 0.596277.

CONCLUSION

In this research, Fuzzy C-Means algorithm works well in clustering customers. The algorithm categorizes customers into three clusters (golden, silver and bronze) with a cluster accuracy rate of 0.596277, which means the accuracy rate is quite good.

REFERENCES


