Clustering Customer Data Using Fuzzy C-Means Algorithm

Nurfaizah 1,*, Fathuzaen 2

1 Information Systems, Faculty of Computer Science; Universitas Amikom Purwokerto; Jl. Let Jen Pol Soemarto, Purwanegara, Purwokerto Utara, Banyumas; e-mail: nurfaizah@amikompurwokerto.ac.id

2 Informatics, Faculty of Computer Science; Universitas Amikom Purwokerto; Jl. Let Jen Pol Soemarto, Purwanegara, Purwokerto Utara, Banyumas; e-mail: zylancer@gmail.com

* Correspondence Author: e-mail: nurfaizah@amikompurwokerto.ac.id

Received: 4 January 2021; Revised: 14 January 2021; Accepted: 12 February 2021; Available online: 24 March 2021

Abstract

The pattern of the service industry is influenced mostly by economic growth. When economic growth rises, the economic activity will also grow as in the case of insurance activities. One of the assets owned by an insurance company is the customer, hence the existence of a loyal or potential customer should be maintained by the insurance company. This study focuses on clustering or grouping the existing customer data in insurance companies using the Fuzzy C-Means (FCM) algorithm. This study uses data from the company for analysis and the results can be used as a basis for insurance companies in making decisions, especially those related to further insurance marketing to customers who have participated in insurance or who are still actively registered in payment insurance. Fuzzy C-Means can be used for clustering the customer...
datasets. It obtained 3 clustering results using Partition Coefficient (PC) in determining the validity index and the centers value was ranged from 0.5 to 1.0.

Keywords: clustering, customers, fuzzy c-means, insurance.

1. Introduction

According to the Indonesian insurance law no. 40 year 2014, insurance is an agreement between two parties, namely the insurance company and the policyholder, which forms the basis for receiving premiums by the insurance company as compensation for providing compensation to the insured or policyholder for loss, damage, costs incurred, loss of profits, or legal liability to a third party that the insured or policyholder may suffer due to an uncertain event or provide payments based on the death of the insured or payments based on the life of the insured with benefits of a predetermined amount and/or based on the results of fund management.

Another kind of insurance system, the Sharia Insurance is a collection of agreements, which consists of an agreement between a sharia insurance company and a policyholder and an agreement between policyholders with managing contributions based on sharia principles to help and protect each other by providing compensation to participants or policyholders for losses, damage, costs incurred, loss of profits, or legal liability to third parties that may be suffered by participants or policyholders due to an uncertain event or providing payments based on the death of the participant or payments based on the participant’s life with benefits of which the amount has been determined and/or based on the results of fund management (Indonesian insurance law no. 40 year 2014).

The operation of an insurance company cannot be separated from the debtor or insurance customer which is a very important asset in a financial institution or bank. The insurance company pays attention to customer profiles because it is a very important part of bank documentation and must be ready if needed at any time. This is the basis for prospective customers to fill out an identity form that contains the condition of an insurance customer.
The customer profile is used as a company reference to find out more about work, personal data, family data. Completeness of customer data makes it easier for the insurance company to collect data and further the insurance company can be the basis for determining or classifying potential and non-potential customers.

The c-means algorithm is one of the algorithms used to cluster data. This method can classify incomplete data using the maximum likelihood criteria and clustering results can provide decision support for companies to increase the effectiveness of raw material storage (Zhang et al., 2016). In addition, C-Means can be used for data processing (Agustini, 2019).

This study focuses on grouping or clustering insurance customer data using the C-Means algorithm. Previous research grouped and produced optimal results in clustering food products in restaurants using the C-Means algorithm (Agustini, 2019). Other research conducted by Rahakabhw implemented Fuzzy C-Means in determining scholarships by grouping data of students who are eligible to receive scholarships and students who are not eligible for scholarships. Fuzzy C-Means can also be used as a method and as a determinant of eligibility for scholarship acceptance (Rahakbauw et al., 2017).

Similar research was also conducted by Rustiyan who implemented the Fuzzy C-Means algorithm to see the problem of mandatory savings for cooperative members and obtained the results of grouping the cooperative customers for each district/city where they live in the payment of Mandatory Savings (Rustiyan R & Mustakim, 2018). Another research theme using the C-Means method was also conducted by Rismanto (Rismanto et al., 2017) which produced student grouping data based on the number of lecture attendance levels for follow-up actions and determined for students to supervise or not.

Various studies related to using the C-Means algorithm have been conducted by previous researchers. The focus of this study is the implementation of the C-Means algorithm for clustering insurance customer profile data. Therefore, the potential and non-potential customer data groupings can be obtained and beneficial as well to other types of insurance in another insurance company.
2. Research Methods
2.1. Study of literature

Clustering

Clustering is a technique that is well known and widely used to group data or objects into groups of data based on similar characteristics (Han et al., 2012; Sari & Suranti, 2016). Groups can be seen as subsets of data sets, one of the possible classification methods of grouping can be adjusted whether the part is fuzzy. The classical grouping method (hard clustering) is based on classical set theory, which determines that an object can be a member or not a member of Fuzzy groupers allow an object to be a member of several groups at once with different degrees of membership. The degrees of membership are between the ranges 0 and 1. Thus, dataset X can be partitioned into c fuzzy parts.

Data clustering is an unsupervised data mining method. There are two types of data clustering that are often used in data grouping processes, namely hierarchical (hierarchical) data clustering and non-hierarchical data clustering. Clustering is a way of entering an observed pattern into an unknown pattern class and is called a pattern cluster Fuzzy clustering is a technique to determine the optimal cluster in a vector space based on the normal Euclidean form for the distance between vectors. Fuzzy clustering is very useful for fuzzy modeling, especially in identifying fuzzy rules.

There are several data clustering algorithms, one of which is Fuzzy Clustering C-Means. The basic concept of FCM is, first, to determine the center of the cluster, which will mark the average location for each cluster. By repeating the cluster center and degree of membership of each data point, it can be seen that the cluster center will move towards the right location. This iteration is based on minimizing the objective function describing the distance from a given data point to the center of the cluster which is weighted by the degree of membership of the data point (Kusumadewi & Purnomo, 2004). The Fuzzy C-Means algorithm is explained as follows:
a. Input data to be cluster X, in the form of a matrix measuring n × m (n = number of sample data, m = attributes of each data). Xij = ith sample data (i = 1, 2, …, n), jth attribute (j = 1, 2, …, m).

b. Determine:
   1) Number of clusters = c
   2) Power = w
   3) Maximum iteration = MaxIter
   4) The smallest error expected = ξ
   5) The initial objective function = Po = 0
   6) Initial iteration = t

c. Generate random values μik, i = 1, 2, …, n; k = 1, 2, …, c as the elements of the initial partition matrix μik. μik is the degree of membership which refers to how likely it is that a data group can become a member of a cluster. The matrix positions and values are constructed randomly. Where membership value is located at intervals of 0 to 1. At the initial position of the partition U matrix is still not accurate as well as the center of the cluster. So that the tendency of data to enter a cluster is also not accurate.

d. Calculate the center of the k-th cluster: Vkj, where k = 1, 2, …, c and j = 1, 2, …, m. where Xij is the fuzzy variable used and w is the weight, as in the following cluster determination formula.

\[
V_{kj} = \frac{\sum_{i=1}^{n} (\mu_{ik})^{w} X_{ij}}{\sum_{i=1}^{n} (\mu_{ik})^{w}}
\]

(1)

e. The objective function is used as a looping condition to get the right cluster center. So that we get the data tendency to enter which cluster at the final step.

f. Calculate the objective function in the t-iteration Pt with the following formula

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

(2)
g. Calculation of the objective function of Pt where the value of the fuzzy Xij variable is less than the center of the cluster Vkj, then the results of the subtraction are squared, then each quadrad result is added up to be multiplied by the quadrad of the degree of membership μik for each cluster. After that, add up all the values in all clusters to get the objective function of Pt.

h. Calculate the change in the partition matrix using the following formula:

\[
\mu_{ik} = \left( \sum_{j=1}^{m} \left( X_{ij} - V_{kj} \right)^2 \right)^{-\frac{1}{w-1}} \left( \sum_{k=1}^{c} \sum_{j=1}^{m} \left( X_{ij} - V_{kj} \right)^2 \right)^{-\frac{1}{w-1}}
\]

where: i = 1,2,... n and k = 1,2, .. c.

To find the change in the μik partition matrix, the reduction in the value of the fuzzy Xij variable is done back to the center of the cluster Vkj and then it is squared. Then add up and then raise the power to -1 / (w-1) with weight, w = 2, the result is that each data is raised to the power of -1. After the calculation process is carried out, normalize all new membership degree data by adding up the new membership degrees k = 1,... c, the result is then divided by the new membership degrees. This process is carried out so that the new membership degree ranges between 0 and not more than 1 and then check the stop condition, if : ( | Pt - Pt-1 | <ξ) or (t> maxIter) then stop the process, otherwise: t = t + 1, and repeat step 4.

2.2. Research methods

The research stage is needed as a framework and guide for the research process, so that a series of research processes can be carried out in a directed, orderly and systematic manner. The research design refers to the six stages of the Cross Industry Standard Process for Data Mining (CRISP-DM) or the life cycle of developing data mining as a framework for data mining projects.
(Fadillah, 2015; Shearer, 2000). The CRISP-DM stage used can be as shown in Figure 1.

In more detail, the CRISP-DM research stage is explained as follows:

a. Business Understanding
   This stage is the stage in identifying business processes and objectives that are used as a basis for determining the patterns to be looked for in the data mining process more specifically related to data clustering using C-Means.

b. Data Understanding
   After identifying the business process, the next stage is understanding the data needs related to the achievement of the objectives in the previously determined research. The process undertaken is to understand the data and retrieve some of the data needed in this study. The data to be used is insurance customer data.

c. Preparation Data
   This stage is the stage where the customer profile data collection process is converted into a data set which will then be processed in pre-processing data.

d. Modeling
After the data goes through the pre-processing and normalization stages, in this stage the modeling stage is carried out using the C-Means algorithm.

e. Evaluation
The results of clustering that have been processed in the modeling stage are then tested for the validity level to determine the optimal number of clusters.

f. Deployment
This stage is the stage of applying the model that has been generated and compiling a clustering report.

3. Results and Discussion

3.1. Business Understanding

Insurance company "XYZ" is an insurance company that has products by looking at the amount of premium that customers take. The existing products include the Prosperous Investment Fund, the Healthy and Prosperous Fund, and the Prosperous Investment Fund. This study will focus on clustering using the FCM algorithm. The understanding of the business carried out in this study looks at the tendency of customers to take products and the amount of premiums.

a. Determining business objectives is done by identifying customer data patterns on products and premium deposits in a certain period.

b. Assessing the situation, the insurance data management carried out by each unit is then put together in one report per company month.

c. Determining the purpose of data mining, this study aims to see the patterns of customers on product selection and grouping the tendency of the premiums taken by customers.

3.2. Data Understanding

The data to be used is insurance customer data taken from customer report data for the entire unit in the 2019 period. Data collection and data selection determination are also based on interviews conducted. The data used is new customer data for the month, the dataset used is 301 customer data. Table 1 below is an initial dataset table that has been manually coded for policy number data and product types using Excel.
Clustering Customer Data Using Fuzzy C-Means Algorithm

Table 1. Initial Dataset

| NO POLICY | PRODUCT | BASIC PREMIUM | TOP UP PREMIUM | GROSS PREMIUM | ANP | FEE BASED (E) |
|-----------|---------|---------------|----------------|---------------|-----|---------------|
| 1         | 1       | 6,000,123     | 6,000,123      | 0             | 6,000,123 | 3,000,062     |
| 2         | 1       | 350,000       | 200,000        | 150,000       | 350,000  | 4,200,000     |
| 3         | 1       | 300,000       | 230,000        | 70,000        | 300,000  | 3,600,000     |
| 4         | 1       | 500,000       | 300,000        | 200,000       | 500,000  | 6,000,000     |
| 5         | 1       | 300,000       | 230,000        | 70,000        | 300,000  | 3,600,000     |
| 6         | 1       | 350,000       | 200,000        | 150,000       | 350,000  | 4,200,000     |
| 7         | 1       | 500,000       | 300,000        | 200,000       | 500,000  | 6,000,000     |
| 8         | 1       | 6,000,000     | 3,000,000      | 3,000,000     | 6,000,000| 6,000,000     |
| 9         | 1       | 10,000,000    | 10,000,000     | 0             | 10,000,000| 5,000,000     |

Source: Research Result (2021)

3.3. Data Preparation

The dataset that has been owned is then changed the format from excel to csv data file in order to simplify the data processing process in the program. As seen in table 2 below is a conversion table from an excel file to a csv file data.

Tabel 2. CSV File Conversion

| NO POLICY, "PRODUCT", "PREMIUM", "BASIC PREMIUM", "TOP UP PREMIUM", "GROSS PREMIUM", "ANP", "FEE BASED (E)" |
|---------------------------------------------------------------|
| 1,"1","6,000,123","6,000,123","0","6,000,123","3,000,062" |
| 2,"1","350,000","200,000","150,000","350,000","4,200,000","101,875" |
| 3,"1","300,000","230,000","70,000","300,000","3,600,000","115,875" |
| 4,"1","500,000","300,000","200,000","500,000","6,000,000","152,500" |
| 5,"1","300,000","230,000","70,000","300,000","3,600,000","115,875" |
| 6,"1","350,000","200,000","150,000","350,000","4,200,000","101,875" |
| 7,"1","500,000","300,000","200,000","500,000","6,000,000","152,500" |
| 8,"1","6,000,000","3,000,000","3,000,000","6,000,000","6,000,000","1,537,500" |
| 9,"1","10,000,000","10,000,000","0","10,000,000","5,000,000" |
| 300,"1","13,000,123","10,000,123","3,000,000","13,000,123","5,037,562" |

Source: Research Result (2021)

3.4. Modeling

After the data goes through the pre-processing stage, the data normalization stage is carried out through the python program as in Figure 2.
below is a way to normalize it by removing the comma and casting it to the float data type.

```python
    data = data[features]
    data = data.replace(',', '', regex=True)

    for key in features:
        data[key] = pd.to_numeric(data[key], downcast="float")
    data.head()
```

Source: Research Result (2021)

Figure 2. Data Type Normalization Process

Table 3 below is the initial display when data is inputted into the program.

One of the results of the normalization process is shown in table 3 below

| Gross Premium | Fee Based (E) |
|---------------|--------------|
| 0             | 6000123.0    | 3000062.0 |
| 1             | 350000.0     | 101875.0  |
| 2             | 300000.0     | 115875.0  |
| 3             | 500000.0     | 152500.0  |
| 4             | 300000.0     | 115875.0  |

Source: Research Result (2021)

Table 3. Results of Data Type Normalization

The next stage is data modeling which is done by inputting the CSV dataset into the program and carrying out the normalization stage. Figure 3 below is a step in modeling the FCM algorithm. The number of clusters is determined as 3 clusters with a maximum interaction of 100.

```python
import skfuzzy as fuzz
import numpy as np

DEF_CLUSTER = 3

# Extract Data to Numpy Array
x = data[features[0]]
y = data[features[1]]
target = np.vstack((x, y))

center, u, u0, d, jm, p, fpc = fuzz.cluster.cmeans(
    target, DEF_CLUSTER, 2, error=0.005, maxiter=1000, init=None
)

# Membership berarti dimana cluster data dari x dan y
# Grup tersebut berupa index dari 0 ke banyak cluster
membership = np.argmax(u, axis=0)
```

Source: Research Result (2021)

Figure 3. Clustering Model
3.5. Evaluation

Based on the validity test of the interactions carried out, the data clustering from the dataset used is shown in Figure 4. There are 3 clusters, i.e. cluster 0, cluster 1 and cluster 3 with 3 cluster center points, namely at 0.5 to 1.0 point.

![Figure 4. Center Cluster](image)

Source: Research Result (2021)

The data processing process is also carried out by assessing the validity index using the Partition Coefficient (PC) to measure the amount of overlapping between the clusters formed and the number of the best clusters aimed at the highest PC value (Haqiqi & Kurniawan, 2015). Figure 5 shows validity index results.

The center having highest PC value was center 2 with a validity value of 0.97. A value close to 1 indicates that the quality of the clusters formed is getting better (Astria & Suprayogi, 2017).

![Figure 5. Validity Index Results](image)

Source: Research Result (2021)
3.6. Deployment

The clustering results show the grouping of members based on the cluster center formed and Fuzzy Clustering means grouping based on the degree of membership. Based on the validity test of the interactions carried out, the data clustering from the dataset used is obtained as follows in Figure 6. There are 3 clusters, namely cluster 0, cluster 1 and cluster 3 with 3 cluster center points, namely at point 0.5 to 1.0.

![Figure 6 Clustering View with Single Plot](image)

Source: Research Result (2021)

Figure 6 Clustering View with Single Plot

The cluster formed based on the dataset used is cluster 0, which is a cluster that has a gross premium of 100,000 to 200,000, while cluster 1 is a cluster that has a gross premium $> 20,000,000$, and cluster 2 is a cluster that has a gross premium $> 200,000$ to 15,000,000.

4. Conclusion

The Fuzzy Clustering Means (FCM) algorithm can be used to cluster insurance customer data. The clustering produced 3 clusters from 300 datasets which were grouped based on the degree of membership of each data, while the clusters formed are cluster 0, cluster 1 and cluster 3 with a cluster center value of 0.5 to 1 which is obtained from the highest validation index and is close to number 1.
Clustering Customer Data Using Fuzzy C-Means Algorithm

Acknowledgments
The authors would like to thank Universitas AMIKOM Purwokerto for providing research support and funding.

Author Contributions
Nurfaizah proposed the topic; Nurfaizah and Fathuzaen conceived models and designed the experiments; Nurfaizah and Fathuzaen the algorithms; Nurfaizah and Fathuzaen analysed the result.

Conflicts of Interest
The author declare no conflict of interest.

References
Agustini, F. (2019). Implementasi Algoritma Fuzzy C-Means Studi Kasus Penjualan di Sushigroove Restaurant. Jurnal Ilmu Pengetahuan Dan Teknologi Komputer, 3(1), 127–132.

Astria, D., & Suprayogi. (2017). Penerapan Algoritma Fuzzy C-Means Untuk Clustering Pelanggan Pada CV. Mataram Jaya Bawen. Eksplora Informatika, 6(2), 169–178.

Fadillah, A. P. (2015). Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ). Jurnal Teknik Informatika Dan Sistem Informasi (JuTISI), 1(3), 260–270. https://doi.org/10.28932/jutisi.v1i3.406

Han, J., Kamber, M., & Pei, J. (2012). Data Mining Concepts and Techniques. Elsevier Inc.

Haqiqi, N. B., & Kurniawan, R. (2015). Analisis Perbandingan Metode Fuzzy C-Means dan Subtractive Fuzzy C-Means. Media Statistika, 8(2), 59–67.

Kusumadewi, S., & Purnomo, H. (2004). Aplikasi Logika Fuzzy untuk Pendukung Keputusan. Graha Ilmu.

Rahakbauw, D. L., Ilwaru, V. Y. I., & Hahury, M. H. (2017). Implementasi Fuzzy C-Means Clustering Dalam Penentuan Beasiswa. Jurnal Ilmu Matematika Dan Terapan, 11(1), 1–12.
Rismanto, R., Fahrur Rozi, I., & Prasetyo, A. (2017). Implementasi Fuzzy C-Means Untuk Prediksi Perilaku Mahasiswa Berdasarkan Jumlah Ketidakhadiran. *SMARTICS Journal*, 3(2), 39–45. https://doi.org/10.21067/smartics.v3i2.1965

Rustiyan R., & Mustakim. (2018). Penerapan Algoritma Fuzzy C Means untuk Analisis Permasalahan Simpanan Wajib Anggota Koperasi. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, 5(2), 171. https://doi.org/10.25126/jtiik.201852605

Sari, H. L., & Suranti, D. (2016). Perbandingan Algoritma Fuzzy C-Means (FCM) dan Algoritma Mixture Dalam Penclusteran Data Curah Hujan Kota Bengkulu. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 7–15. https://journal.uit.ac.id/Snati/article/viewFile/6228/5589

Shearer. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5(4).

Zhang, L., Lu, W., Liu, X., Pedrycz, W., & Zhong, C. (2016). Fuzzy C-Means Clustering of Incomplete Data Based on Probabilistic Information Granules of Missing Values. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2016.01.048