(IJID) International Journal on Informatics for Development,
Vol. 3, No. 1, 2014

Patient Data Clustering using Fuzzy C-Means (FCM) and Agglomerative Hierarchical Clustering (AHC)

Rosalia Susilowati
Informatics Department
Islamic State University (UIN) of
Sunan Kalijaga
Yogyakarta, Indonesia

Ahmad Subhan Yazid
Informatics Department
Islamic State University (UIN) of
Sunan Kalijaga
Yogyakarta, Indonesia
yazid.anfalah@gmail.com

Shofwatul ‘Uyun
Informatics Department
Islamic State University (UIN) of
Sunan Kalijaga
Yogyakarta, Indonesia
Shofwatul.uyun@uin-suka.ac.id

***Abstract*—Generally, the current system development only include the input, view, and reports. At Jogja Hospital, a system with a patient database can only provide information about the percentage of male and female patients. Its unable to extract more specific information, even though medical record data has a lot of information. The complete information should be used as a reference for the authorities to make a decision. This information can be obtained by analyzing and processing the medical record data. One way to extract information from this data is clustering. The domain of this study is patient data. Before the data is clustered, preprocessing is needed through name standardization, numeration, and data normalization. During the clustering process, the algorithms used are Fuzzy C-Means (FCM) and Agglomerative Hierarchical Clustering (AHC). Two algorithms are implemented to determine which algorithm is the most appropriate and fast to handle the processing of patient data. The results of the study show that the processing time required to do clustering with FCM algorithm is relatively faster than AHC algorithm. For data with small volumes, the iteration of FCM algorithm is more than AHC algorithm, however, the results of the clustering using FCM algorithm are easier to interpret than AHC algorithm. From the visualization of clustering results, found that the cluster pattern with FCM algorithm is better based on the three variables used as references. So the most suitable algorithm to use is Fuzzy C-Means (FCM) for processing patient data.**

Keywords-Agglomerative Hierarchical Clustering; Clustering; Fuzzy C-Means; Patient data

I. Introduction

The need for information in human life has become a basic need, including in education, social, political, cultural or health. This will be a problem when these needs cannot be formulated properly so that organizational goals are not met. For example in a hospital environment, information about an illness or a patient who has a particular disease is very important. This information can be obtained from medical record data.

Medical records are files that contain notes and documents about the patient's identity, history of examination, treatment, actions, and other services to the patient when the patient conducts a health check on health care facilities [1]. From medical record data, a lot of information can be obtained. This information can be used as a consideration for the hospital to take action to prevent or overcome it. This information can be obtained from processing and analyzing the patient's medical record data.

Jogja Hospital is one of the public hospitals that have a large number of patients with diverse backgrounds and diseases. In this hospital, there are eleven poly ready to serve patients that are internal, surgical, child, neurologist, mental, *THT*, eye, skin and genital, dental and oral and other services or checks up. The results of the year-end report on patient visits from 2007-2011 noted that the services most frequently visited by patients were emergency care, outpatient and pioneer care while the percentage of diseases diagnosed was: 31% internal disease, 24% child, 10.9% obsgin, surgery 13.3%, nerves 8.7%, eyes 1.4%, ENT 0.4% and skin 0.3%.

With abundant data, it is necessary to have a data processing management system. During this time, the year-end report made by the hospital was limited to the percentage of men and women and the percentage of diagnoses suffered by patients. The report has not been able to analyze patient data based on age, gender, address, and other criteria. From these conditions, the hospital needs to know information about patient visits and the patient's tendency based on address, gender, and age to be used as a support for decision making for the hospital and related agencies in planning the future strategy.

To be able to help in the process of finding information in large data, the data mining technique can help this process. Data mining is a concept of searching for information from very large data sets [2]. Data mining is intended to provide real solutions for decision makers. Broadly speaking, the objectives of data mining are divided into two parts, namely descriptive and predictive [3]. Descriptive that is describing everything that has happened, the techniques included in this category are association and clustering [4]. Predictive is predicting everything that will happen. The techniques included in this category are classification (classification) and function approximation (function approach) [5].

Data mining techniques used in this study are clustering based on the analysis of patient data obtained from the Jogja Hospital and based on the objectives to be obtained, namely the classification of patients. In this case, there are two clustering algorithms used, namely Fuzzy C-Means and Agglomerative Hierarchical Clustering (AHC) Algorithms. FCM determines the optimal cluster in a vector space based on the normal euclidian form to assess the distance between vectors [6]. AHC combines two clusters with the closest distance until finally a cluster consists of the entire cluster is formed [7]. The selection of the two algorithms aims to compare the reliability of the two in producing the most approMalete clustering results.

II. Methods

This study was conducted at Jogja Hospital located in Wirosaban, Yogyakarta. The object of this study is patient data of that hospital.

A. Data and Data Collection

There are two types of data, namely primary data and secondary data. Primary data is data taken directly by researchers using research instruments. Primary data obtained through interviews and direct observation of patient data. From patient data taken are fields of age, gender, and address. Secondary data is supporting research data obtained from the year-end report of Jogja Hospital.

B. Research Flow

The flow of research from clustering patient data using the Fuzzy C-Means algorithm and Agglomerative Hierarchical Clustering is as follows:

1) Preliminary Study

In this stage, a preliminary study is conducted to find out the problems and opportunities of making patient data clustering and asks for direction from management to find out the scope and feasibility of the application design to be made.

2) Needs Analysis

Defining information needs for the patient data clustering process.

3) Designing a System Prototype

The development of this system uses the basis of data mining. There are several stages of data mining. Of the seven steps of data mining [4] will be included in the system design process which includes:

a) Data Input

The inputted data is patient data in the .xls format. In the input file, there are only age, Gender and address fields.

b) Preprocessing

The preprocessing stage of the system involves three stages in data mining, namely data cleaning, data selection, and data transformation. Each of these processes can be explained as data cleaning, data selection, data transformation, and Normalization

c) Clustering

The algorithms used in the clustering process are FCM and AHC. In this stage, clustering steps are carried out with fuzzy c-means and agglomerative hierarchical clustering AHC based on age, gender and patient's address.

d) Cluster Pattern Output

To produce cluster patterns, the stages of data mining used are:

- Pattern evaluation: To identify interesting patterns into knowledge-based found.
- Knowledge Representation: a visualization and presentation of knowledge about the methods used to obtain knowledge gained by users, visualization in images of cluster patterns.

e) Make a System Prototype

Translating design results in computer programs using Borland Delphi programming language and database using MySQL.

In general, the flow of the system can be seen in Figure 1.

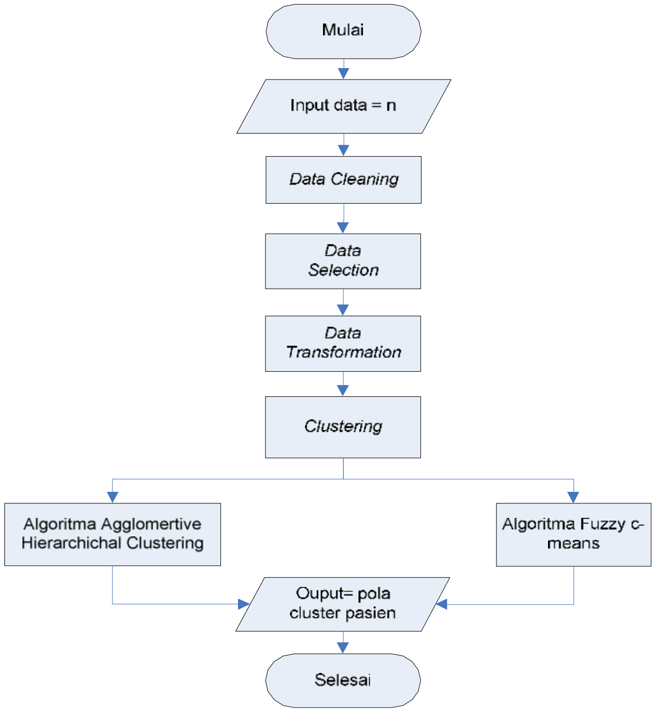


Figure 1. System Outline Flow Chart.

III. Implementation

Before the data is extracted, there are stages of preparing data. The existing data must be processed through the preprocessing stage for the next clustering process. In summary, the flow of the system can be seen in Figure 2.

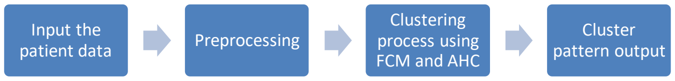


Figure 2. Stages of system processes

A. Data Input

The data used is patient data which is still in the .xls format because there is no database for the clustering process. The data available in the patient archive includes the following attributes: Entry date - Exit date - Exit method - Name - No Medical Record - Gender - Patient's age - Patient's address - Pay status - Ward - Origin of referral - Diagnosis

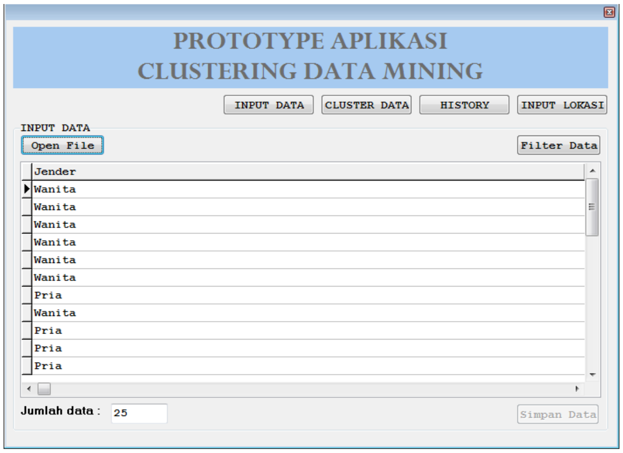
Not all of the attributes are included in the data mining process, only the attributes that support the clustering process are used, among others: Gender (which will be changed to Gender), Age of the patient (to be changed to age), and Address of the patient (to be changed to address). Table I is an example of raw data that will be processed in the application.

TABLE I. Sample Data

Gender	Age	Address
Female	0 Y 0 M 0 D	Badran Jt I/1026
Female	63 Y 0 M 14 D	Krapyak Wetan Rt 05 Pangungharjo Sewon Bantul
Female	0 Y 0 M 0 D	Banyakan Iii Rt02 Piyungan Bantul
Female	54 Y 3 M 23 D	Keparakan Lor Mg I/834 Rt 42 Rw 09 Keparakan Mergangsan Yogyakarta
Female	0 Y 0 M 1 D	Blarangan Rt 01/04 Gk
Female	0 Y 0 M 1 D	Blarangan Rt 01/04 Gk
Male	62 Y 9 M 13 D	Pringgokusuman Gt Ii/2133 Pringgokusuman Gedongtengen
Female	10 Y 0 M 25 D	Keparakan Lor Mg I/696 Rt 36/08 Keparakan Mergangsan Yogyakarta
Male	68 Y 0 M 24 D	Minggiran Mj 2/ 1291
Male	23 Y 1 M 8 D	Semaki Kulon Uh I/311
Male	23 Y 1 M 8 D	Semaki Kulon Uh I/311
Male	21 Y 11 M 23 D	Keparakan Lor Mgi/874 Rt43/09
Male	65 Y 4 M 9 D	Purbayan Rt 52 Kotagede Yogyakarta
Male	0 Y 0 M 0 D	Pringgokusuman No.26
Female	72 Y 0 M 14 D	Perum Jatimulyo Baru Rt26
Female	2 Y 6 M 18 D	Rusunawa Panggungharjo / Druwo Rt.02 Bangunharjo Sewon Bantul
Female	51 Y 9 M 14 D	Dongkelan Rt.06/00 Kasihan
Male	71 Y 1 M 6 D	Kweni Rt 07 No 403 Panggungharjo Sewon Abntul
Female	44 Y 2 M 24 D	Keparakan Kidul Rt 56/13 Mg I/1212 Yogyakarta
Male	73 Y 9 M 20 D	Munggur Rt 03/15 Yk
Female	52 Y 5 M 23 D	Gambiran Rt 44/11 No 374 Pandeyan Umbulharjo Yk
Male	0 Y 8 M 22 D	Rejokusuman Rt 03 Rw 07 Tamanan Banguntapan Bantul
Female	26 Y 4 M 16 D	Karang Jambe Rt 04/19 Banguntapan Bantul
Male	54 Y 3 M 29 D	Karanganom Rt 01 Bantul
Female	35 Y 8 M	Krambil Sawit Rt.06/02 Pringwulung, Saptosari,

	18 D	Gunungkidul

In the application, the data input process can be seen in Figure 3. Users simply press the "Open File" button and then select the data to be clustered.



Jender
Wanita
Wanita
Wanita
Wanita
Wanita
Wanita
Pria
Wanita
Pria
Pria
Pria

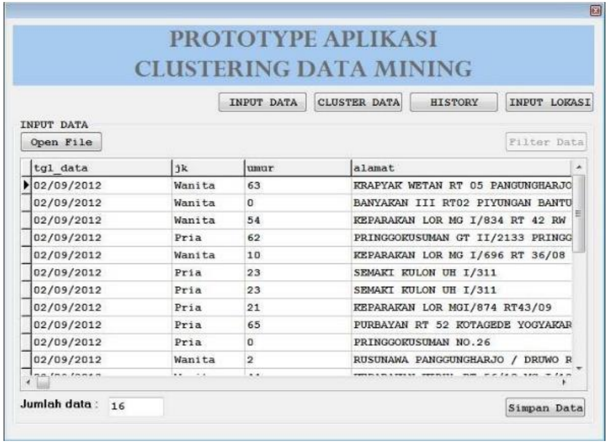
Jumlah data : 25

Figure 3. Data Input Process

From all data in Table I, it will be filtered based on patients having addresses in each village in the Yogyakarta City area. The results are as in Table II. After filtering, only 10 data from the initial data will be processed for clustering as shown in Figure 4.

TABLE II. FILTERING RESULT

No	Gender	age	Address
1	Female	54 Y 3 M 23 D	Keparakan Lor Mg I/834 Rt 42 Rw 09 Keparakan Mergangsan Yogyakarta
2	Male	62 Y 9 M 13 D	Pringgokusuman Gt Ii/2133 Pringgokusuman Gedongtengen
3	Female	10 Y 0 M 25 D	Keparakan Lor Mg I/696 Rt 36/08 Keparakan Mergangsan Yogyakarta
4	Male	23 Y 1 M 8 D	Semaki Kulon Uh I/311
5	Male	23 Y 1 M 8 D	Semaki Kulon Uh I/311
6	Male	21 Y 11 M 23D	Keparakan Lor Mgi/874 Rt43/09
7	Male	65 Y 4 M 9 D	Purbayan Rt 52 Kotagede Yogyakarta
8	Male	0 Y 0 M 0 D	Pringgokusuman No.26
9	Female	44 Y 2 M 24 D	Keparakan Kidul Rt 56/13 Mg I/1212 Yogyakarta
10	Female	52 Y 5 M 23 D	Gambiran Rt 44/11 No 374 Pandeyan Umbulharjo Yk



tgl_data	jk	umur	alamat
02/09/2012	Wanita	63	KRAPYAK WETAN RT 05 PANGUNGHARJO
02/09/2012	Wanita	0	BANYAKAN III RT02 PIYUNGAN BANTU
02/09/2012	Wanita	54	KEPARAKAN LOR MG I/834 RT 42 RW
02/09/2012	Pria	62	PRINGGOKUSUMAN GT II/2133 PRINGG
02/09/2012	Wanita	10	KEPARAKAN LOR MG I/696 RT 36/08
02/09/2012	Pria	23	SEMAKI KULON UH I/311
02/09/2012	Pria	23	SEMAKI KULON UH I/311
02/09/2012	Pria	21	KEPARAKAN LOR MGI/874 RT43/09
02/09/2012	Pria	65	PURBAYAN RT 52 KOTAGEDE YOGYAKAR
02/09/2012	Pria	0	PRINGGOKUSUMAN NO.26
02/09/2012	Wanita	2	RUSUNAWA PANGGUNGHARJO / DRUWO R

Jumlah data : 16

Figure 4. Filtered Data

After filtering, the data can be saved by pressing the "Save Data" button. This stored data is master data which will then be preprocessed.

B. Preprocessing

Of the three attributes used in the data mining process, it is necessary to go through a process of cleaning and changing forms. The process of preparing and refining data through the following steps:

1) Data Cleaning

Data cleaning is a process of removing noise and inconsistent data or irrelevant data. In the medical record database that contains imperfect entries such as missing data, invalid data or incorrect typing is removed.

2) Standardization of naming (Data Selection)

Data selection is the selection or selection of data that is suitable for analysis to be taken from the database. Not all data in the medical record database is used for the clustering process. The data used are only patient identity data, namely age, gender, and address.

This step is done regarding the patient's address. The area covered is only the city of Yogyakarta which is divided into 46 urban villages. So the patient data will be sorted by a string containing the name of the village in Yogyakarta. Furthermore, the age of patient data included the number of years and months. In this study only the year was taken. For example, 3 years 5 months, then only 3 years will be processed.

3) Data Transformation

Data transformation is the process of converting or merging into a format suitable for processing in data mining. This step is carried out on gender attributes and addresses so it is suitable for data processing in the patient data clustering stage. Changing forms is done by discretization [8]. The attributes that need to be changed are gender and address. The male gender is given a value of 0 and the woman is given a value of 1. Likewise done for the address given a value of 0 for the Mantrijeron and so on until the value of 45.

4) Normalization

Normalization aims to give the same value weight to all different variable data scales [9]. With this standardization, the value of an attribute is normalized based on the average value (m) using equations (1) and standard deviation (s) attribute values with equation (2). After changing the shape, the results of the data and the average (µ) will look like in Table III.

TABLE III. DATA AFTER PREPROCESSING PROCESS

Gender	Age	Address
1	54	7
0	62	43
1	10	7
0	23	33
0	23	33
0	21	7
0	65	42
0	0	43
1	44	7
1	52	35
$\Sigma^n \quad f = 4 \;\; s=1$	$\Sigma^n \; f = 354 \;\; s=1$	$\Sigma^n \; f = 257 \;\; s=1$
μ_f = 0,26666667	μ_f = 23,6	μ_f = 17,13333

The next step is to calculate the Mean Absolute Deviation (Sf) value based on each data set, age, and address. The equation used is

$$S_f = 1/n\,(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots\ |x_{nf} - m_f|) \qquad (1)$$

Where x1f ... xnf is the n-variable of f and mf is the average value of f.

For m = 1 / n (x1f + x2f + ... + xnf) (2)

The result is

$$S_{f1} = \frac{1}{10}(|1- 0{,}266666667| + |0-0{,}266666667| + |1-0{,}266666667| + |0-0{,}266666667| + |0-0{,}266666667| + |0-0{,}266666667| + |0-0{,}266666667| + |0-0{,}266666667| + |0-0{,}266666667| + |1-0{,}266666667| + |1-0{,}266666667|)$$

$$S_{f1} = \frac{1}{10}(0{,}733333 + 0{,}266667 + 0{,}733333 + 0{,}266667 + 0{,}266667 + 0{,}266667 + 0{,}266667 + 0{,}266667 + 0{,}733333 + 0{,}733333)$$

$$S_{f1} = 0{,}302222$$

$$S_{f2} = \frac{1}{10}(|54-23{,}6|+|62-23{,}6|+|10-23{,}6|+|23-23{,}6|+|23-23{,}6|+|21-23{,}6|+|65-23{,}6|+|0-23{,}6|+|44-23{,}6|+|52-23{,}6|)$$

$$S_{f2} = \frac{1}{10}(30{,}4 + 38{,}4 + 13{,}6 + 0{,}6 + 0{,}6 + 2{,}6 + 41{,}4 + 23{,}6 + 20{,}4 + 28{,}4)$$

$$S_{f2} = 13{,}33333$$

$$S_{f3} = \frac{1}{10}(|7-17{,}13333|+|43-17{,}13333|+|7-17{,}13333|+|33-17{,}13333|+|33-17{,}13333|+|7-17{,}13333|+|42-17{,}13333|+|43-17{,}13333|+|7-17{,}13333|+|35-17{,}13333|)$$

$$S_{f3} = \frac{1}{10}(10{,}13333 + 25{,}86667 + 10{,}13333 + 15{,}86667 + 15{,}86667 + 10{,}13333 + 24{,}86667 + 25{,}86667 + 10{,}13333 + 17{,}86667)$$

$$S_{f3} = 11{,}11556$$

The last step is the calculation of the results of normalization using equation 3,

$$z = \frac{s_{1f} - m_f}{c_f} \qquad (3)$$

The results of normalization as shown in Table IV.

TABLE IV. NORMALIZATION RESULT

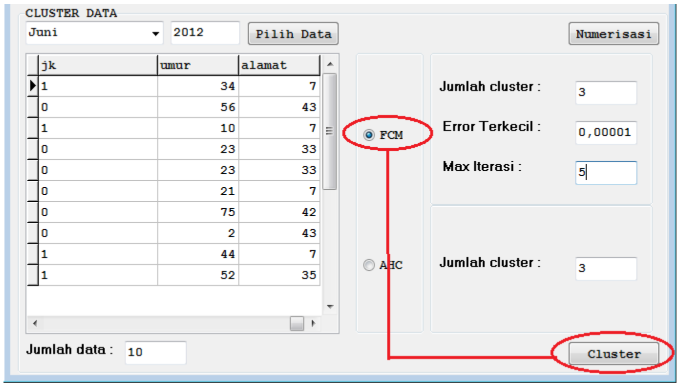
Gender	age	Address
2,426470588	2,28	0,911635346
0,882352941	2,88	2,327069172
2,426470588	1,02	0,911635346
0,882352941	0,045	1,427429028
0,882352941	0,045	1,427429028
0,882352941	0,195	0,911635346
0,882352941	3,105	2,237105158
0,882352941	1,77	2,327069172
2,426470588	1,53	0,911635346
2,426470588	2,13	1,607357057

C. Clustering

The next process is clustering. Users can determine which algorithm will be used.

1) Fuzzy C-Means Algorithm

The first process will use the Fuzzy C-Means algorithm, so the parameters that need to be inputted are "Number of Clusters", "Smallest Error" and "Max Iteration". After the parameters are inputted, then click the "Cluster" button, then the application will do the clustering process (image 5).



CLUSTER DATA

Juni ▾ 2012

jk	umur	alamat
1	34	7
0	56	43
1	10	7
0	23	33
0	23	33
0	21	7
0	75	42
0	2	43
1	44	7
1	52	35

● FCM

Jumlah cluster : 3

Error Terkecil : 0,00001

Max Iterasi : 5

○ AHC

Jumlah cluster : 3

Jumlah data : 10

Figure 5. Selection of FCM Algorithm

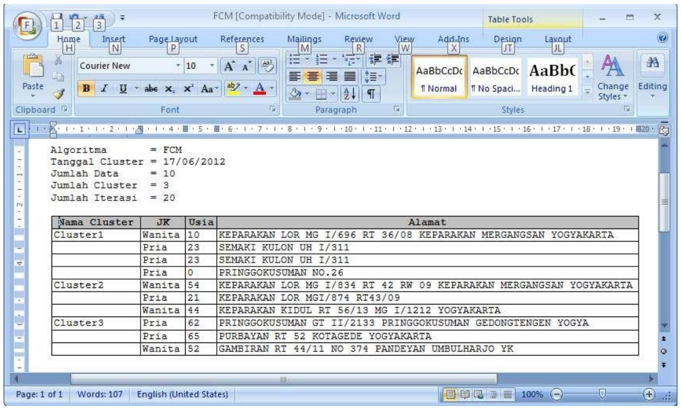
The FCM algorithm accepts input parameters, namely k or the number of clusters, t or the number of iterations and ξ or the smallest error. The initial step of this algorithm is to generate random numbers as the initial partition matrix elements used to normalize the data. Then the center of the cluster will be calculated, for example, there are five clusters, meaning that there will be five cluster centers as well.

The next step is to calculate the objective function. After the objective function is known, the partition matrix will be updated. Then check the stop condition, if (| Pt - Pt-1 | <ξ) or the objective function t minus the objective function t-1 smaller than the smallest error or (t> MaxIter) which is an iteration greater than the maximum iteration then the calculation process stops and a number of clusters are obtained based on the initial input.

With five iterations and 3 clusters, clustering results can be seen in Table V. Clustering is based on the largest degree of membership. If the largest membership degree is located in the first column shows that the data is included in the first cluster, and so on. After going through the calculation process in the application, the results of clustering with the application can be seen in Figure 6.

TABLE V. RESULTS OF CLUSTERING USING THE FCM ALGORITHM

Data			Membership Degree			*Cluster*		
Gender	Age	Address	Gender	age	Address	1	2	3
1	54	7	0,038582	**0,894089**	0,067328		*	
0	62	43	0,008937	0,012910	**0,978151**			*
1	10	7	**0,624805**	0,277376	0,097817	*		
0	23	33	**0,916617**	0,044956	0,038426	*		
0	23	33	**0,916617**	0,044956	0,038426	*		
0	21	7	**0,473298**	0,428562	0,098139	*		
0	65	42	0,016230	0,025431	**0,958338**			*
0	0	43	**0,786889**	0,109536	0,103573	*		
1	44	7	0,000790	**0,998439**	0,000769		*	
1	52	35	0,047515	0,075474	**0,877009**			*



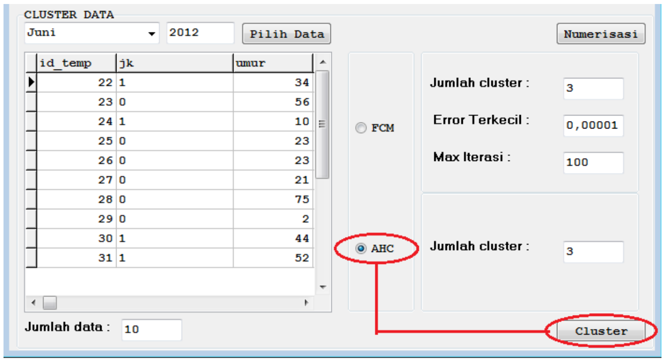
```
Algoritma        = FCM
Tanggal Cluster  = 17/06/2012
Jumlah Data      = 10
Jumlah Cluster   = 3
Jumlah Iterasi   = 20
```

Nama Cluster	JK	Usia	Alamat
Cluster1	Wanita	10	KEPARAKAN LOR MG I/696 RT 36/08 KEPARAKAN MERGANGSAN YOGYAKARTA
	Pria	23	SEMAKI KULON UH I/311
	Pria	23	SEMAKI KULON UH I/311
	Pria	0	PRINGGOKUSUMAN NO.26
Cluster2	Wanita	54	KEPARAKAN LOR MG I/834 RT 42 RW 09 KEPARAKAN MERGANGSAN YOGYAKARTA
	Pria	21	KEPARAKAN LOR MGI/874 RT43/09
	Wanita	44	KEPARAKAN KIDUL RT 56/13 MG I/1212 YOGYAKARTA
Cluster3	Pria	62	PRINGGOKUSUMAN GT II/2133 PRINGGOKUSUMAN GEDONGTENGEN YOGYA
	Pria	65	PURBAYAN RT 52 KOTAGEDE YOGYAKARTA
	Wanita	52	GAMBIRAN RT 44/11 NO 374 PANDEYAN UMBULHARJO YK

Figure 6. Clustering Results with the FCM Algorithm

2) Agglomerative Hierarchical Clustering (AHC) algorithm

To cluster using the AHC algorithm, select the AHC algorithm in the application (Figure 7). Then enter the parameter "Number of Clusters" after that press the "Cluster" button.



Juni 2012

id_temp	jk	umur
22	1	34
23	0	56
24	1	10
25	0	23
26	0	23
27	0	21
28	0	75
29	0	2
30	1	44
31	1	52

○ FCM
Jumlah cluster : 3
Error Terkecil : 0,00001
Max Iterasi : 100

● AHC
Jumlah cluster : 3

Jumlah data : 10

Figure 7. Selection of the AHC Algorithm

AHC algorithm accepts a parameter, k, the number of clusters desired. The first step in the algorithm is to calculate the initial proximity matrix between each pair of points. Thus the initial proximity matrix measures n x n, where n is the number of data points that will decrease by 1 x 1 each iteration. Then the merging steps are performed and the proximity matrix update is performed. The loop will stop after the k cluster is obtained.

The next step is to merge the two closest clusters. The two closest clusters are determined by selecting the smallest value in the proximity matrix. Then take the row and column index which indicates the index of the two clusters. In programs, data in clusters with large indexes will be moved into clusters with small indexes and large cluster indexes will be deleted.

The third step is to renew the proximity of the new cluster proximity matrix. Cluster results are combined in the previous step with another cluster. Program calculations are obtained using the Lance Williams formula with coefficients according to the chosen cluster method. After being combined in a row until it gets k clusters and clustering results.

After normalization with z-score, which produces Table IV, then the proximity matrix is sought by calculating Euclidean Distance from each data. The overall results from the calculation of Euclidean Distance can be seen in Table VI.

TABLE VI. INITIAL EUCLIDEAN DISTANCE

Data	1	2	3	4	5	6	7	8	9	10
1	-	3,207	2,200	3,124	3,124	2,657	3,180	4,173	0,500	1,874
2	3,207	-	4,109	2,061	2,061	3,161	0,164	3,100	3,307	2,208
3	2,200	4,109	-	2,789	2,789	2,154	4,168	3,221	1,700	2,813
4	3,124	2,061	2,789	-	0	1,740	2,184	1,330	2,909	2,541
5	3,124	2,061	2,789	0	-	1,740	2,184	1,330	2,909	2,541
6	2,657	3,161	2,154	1,740	1,740	-	3,211	2,625	2,379	3,200
7	3,180	0,164	4,168	2,184	2,184	3,211	-	3,250	3,303	2,231
8	4,173	3,100	3,221	1,330	1,330	2,625	3,250	-	3,869	3,374
9	0,500	3,307	1,700	2,909	2,909	2,379	3,303	3,869	-	1,913
10	1,874	2,208	2,813	2,541	2,541	3,200	2,231	3,374	1,913	-

From Table VI it is known that the closest distance between two clusters is 0, namely in cluster 4 and cluster 5 so the two clusters are combined. After the merging process, the matrix is repaired with the Lance Williams Function for the Single Linked method with Equation 4.

$$d_{(AB)C} = \text{Min}\ \{d_{AC}, d_{BC}\} \tag{4}$$

After repairing the matrix in the 1st iteration, the new Euclidean Distance can be seen in Table VII.

TABLE VII. EUCLIDEAN DISTANCE IN 1ST ITERATION

	4,5	1	2	3	6	7	8	9	10
4,5	-	3,124	2,061	2,789	1,740	2,184	1,330	2,909	2,541
1	3,124	-	3,207	2,200	2,657	3,180	4,173	0,500	1,874
2	2,061	3,207	-	4,109	3,161	0,164	3,100	3,307	2,208
3	2,789	2,200	4,109	-	2,154	4,168	3,221	1,700	2,813
6	1,740	2,657	3,161	2,154	-	3,211	2,625	2,379	3,200
7	2,184	3,180	0,164	4,168	3,211	-	3,250	3,303	2,231
8	1,330	4,173	3,100	3,221	2,625	3,250	-	3,869	3,374

9	2,909	0,500	3,307	1,700	2,379	3,303	3,869	-	1,913
10	2,541	1,874	2,208	2,813	3,200	2,231	3,374	1,913	-

From Table VII it is known that the closest distance between two clusters is 0, 164220, namely in cluster 2 and cluster 7 that the two clusters are combined. After the merging process, a matrix improvement for the 2nd iteration is performed with the Lance Williams Function for the Single Linked method with Equation 4. In the same way, continue to the seventh iteration.

Of all the iterations, a combination of cluster combinations can be seen in Table VIII.

TABLE VIII. COMBINATION OF CLUSTER MERGERS

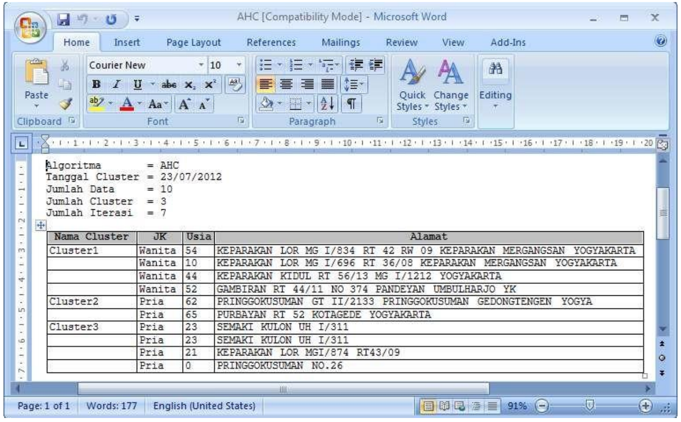
Iteration	*ClusterCombination*		Result
	Cluster 1	*Cluster 2*	
Iterasi ke-1	4	5	4,5
Iterasi ke-2	2	7	2,7
Iterasi ke-3	1	9	1,9
Iterasi ke-4	4	8	4,5,8
Iterasi ke-5	1	3	1,3,9
Iterasi ke-6	4	6	4,5,6,8
Iterasi ke-7	1	10	1,3,9,10

In the 1st iteration, there is a merger of two clusters, cluster 4 and cluster 5 into one new cluster (4,5). In the second iteration of cluster 2 and cluster 7 join into clusters (2,7). In the 3rd iteration merging clusters between cluster 1 and cluster 9 into clusters (1.9). For the 4th iteration, cluster merger occurs between clusters (4,5) and 8 into clusters (4,5,8). Continuing on the fifth iteration occurs a combination of clusters (1.9) and cluster 3 into clusters (1,3.9). In the 6th iteration of clusters (4,5,8) join cluster 6 to form clusters (4,5,6,8). The last iteration is the 7th iteration which occurs between the clusters (1,3,9) and cluster 10 into clusters (1,3,9,10).

After the final iteration, the results of clustering use AHC for 10 data that are clustered into three clusters can be seen in Table IX. The results of clustering in applications with the AHC algorithm can be seen in Figure 8.

TABLE IX. RESULTS OF CLUSTERING USING AHC ALGORITHM

Cluster	Data	Gender	Age	Address
1	1	Female	54	Keparakan Lor Mg I/834 Rt 42 Rw 09 Keparakan Mergangsan Yogyakarta
	3	Female	10	Keparakan Lor Mg I/696 Rt 36/08 Keparakan Mergangsan Yogyakarta
	9	Female	44	Keparakan Kidul Rt 56/13 Mg I/1212 Yogyakarta
	10	Female	52	Gambiran Rt 44/11 No 374 Pandeyan Umbulharjo Yk
2	2	Male	62	Pringgokusuman Gt Ii/2133 Pringgokusuman Gedongtengen Yogya
	7	Male	65	Purbayan Rt 52 Kotagede Yogyakarta
3	4	Male	23	Semaki Kulon Uh I/311
	5	Male	23	Semaki Kulon Uh I/311
	6	Male	21	Keparakan Lor Mgi/874 Rt43/09
	8	Male	0	Pringgokusuman No.26



```
Algoritma        = AHC
Tanggal Cluster  = 23/07/2012
Jumlah Data      = 10
Jumlah Cluster   = 3
Jumlah Iterasi   = 7
```

Nama Cluster	JK	Usia	Alamat
Cluster1	Wanita	54	KEPARAKAN LOR MG I/834 RT 42 RW 09 KEPARAKAN MERGANGSAN YOGYAKARTA
	Wanita	10	KEPARAKAN LOR MG I/696 RT 36/08 KEPARAKAN MERGANGSAN YOGYAKARTA
	Wanita	44	KEPARAKAN KIDUL RT 56/13 MG I/1212 YOGYAKARTA
	Wanita	52	GAMBIRAN RT 44/11 NO 374 PANDEYAN UMBULHARJO YK
Cluster2	Pria	62	PRINGGOKUSUMAN GT II/2133 PRINGGOKUSUMAN GEDONGTENGEN YOGYA
	Pria	65	PURBAYAN RT 52 KOTAGEDE YOGYAKARTA
Cluster3	Pria	23	SEMAKI KULON UH I/311
	Pria	23	SEMAKI KULON UH I/311
	Pria	21	KEPARAKAN LOR MGI/874 RT43/09
	Pria	0	PRINGGOKUSUMAN NO.26

Figure 8. Results of Clustering with AHC Algorithm

IV. RESULT

The experiment was carried out by varying the amount of data to be clustered to find out which algorithm was most appropriate to be used in the case of grouping patients. Variation in the amount of raw data to be clustered starts from 25 data, 50 data, 100 data, 200 data, 500 data, 1000 data, and 1495 data.

The seven experiments that have been carried out include the time needed for the clustering process and the number of iterations. Table X is the result of a summary of recording the time of the clustering process with seven variations in the amount of data clustered. From each graph, the results of clustering show different cluster patterns between the FCM algorithm and the AHC algorithm. The results of clustering with the FCM algorithm are dominated by age and address data. While the AHC algorithm results are more focused on grouping data on gender and age.

TABLE X. RESULTS OF CLUSTERING PATIENT DATA

Raw Data Total	Ready DataTotal	*Cluster*	Processing Time	
			FCM	AHC
25 Data	16 Data	3	4 Detik	5 Detik
50 Data	28Data	3	10 Detik	26 Detik
100 Data	56 Data	5	71 Detik	20 Detik
200 Data	117 Data	5	19 Detik	22 Detik
500 Data	295 Data	10	31 Detik	38 Detik
1000 Data	558 Data	10	1Menit 6 Detik	1 Menit 18 Detik
1495 Data	865 Data	10	1 Menit 34 Detik	1 Menit 50 Detik

The graph of the comparison of the number of iterations between the FCM and AHC algorithms can be seen in Figure 9.

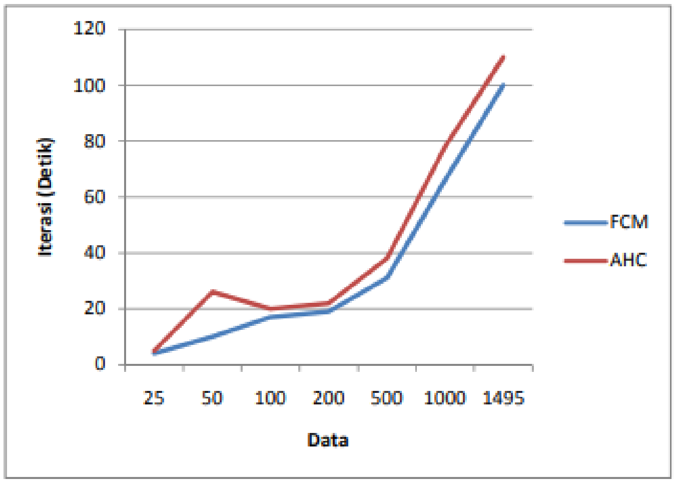


Figure 9. Time Comparison Chart

From the results of the experiment, changes in the volume or amount of data affect the results of clustering. Not only from the volume of data but the results of clustering are also sensitive to changes in the input value of the number of clusters (k). In this experiment an algorithm is chosen which both require input number cluster parameters, to make it easier to analyze the results and compare the resulting cluster patterns.

The clustering results with the FCM algorithm are very affected by the presence of random numbers used for data normalization so when clustering is repeated for the same data, clustering results can be different in membership. Unlike the AHC algorithm, the normalization uses the data itself. So, even if done repeatedly, the results will remain the same.

For the number of iterations, the FCM algorithm can be maximal iterations can be limited, while the AHC iteration algorithm is done is the amount of data minus the number of clusters or (n-k) iterations. Likewise for the time of the clustering process, from the results of the experiments, it can be seen that the time needed for clustering with the AHC algorithm is longer than FCM. This is because of the influence of the iterations carried out. So the FCM algorithm works relatively faster than the AHC algorithm.

In some experiments, especially for relatively small volumes of data, the results of clustering with the AHC and FCM algorithms do not differ greatly in terms of the cluster patterns produced. However, for experiments with large volumes of data, the results of the two algorithms differ greatly. The FCM algorithm that represents partitional clustering produces cluster patterns that are more natural and easily interpreted when compared to the AHC algorithm. Unlike FCM, AHC which is an example of hierarchical clustering only classifies data that is similar in terms of gender so that information from each of the clusters produced is less able to be extracted. When viewed from the results of clustering visualization, the clustering pattern with the FCM algorithm is more grouped based on the three variables, while the results of the AHC algorithm pattern are dominated by Gender data only, so it looks spread and irregular.

V. Conclusion

Jogja Hospital's patient data can be clustered into clusters that have relatively same (homogeneous) properties based on age, gender and address variables. From the clustering results, we can see the trend pattern of patients who go to Jogja Hospital.

The various data domain of patients is quite diverse and the volume is not too large compared to the data warehouse in general, causing the clustering results to be sensitive to changes in the value of parameters and algorithms used. From the experiments conducted in this study, the processing time required to do clustering with FCM algorithm is relatively faster than AHC algorithm. For data with small volumes, the iterations of FCM algorithm are more than AHC algorithm. However, the clustering results of FCM algorithm are easier to interpret than AHC algorithm because AHC algorithm only classifies similar data from Gender variables so less information can be obtained from cluster patterns. When viewed from the visualization of clustering results, the cluster pattern with FCM algorithm is more grouped based on the three variables. So for the patient data domain in this study, the most suitable algorithm to use is Fuzzy C-Means algorithm (FCM.)

References

[1] M. & A. A. Hanafiah, *Etika Kedokteran dan Hukum Kesehatan*. Jakarta: EGC, 2009.

[2] A. & T. C. T. Kadir, *Pengenalan Teknologi Informasi*. Yogyakarta: Andi, 2003.

[3] K. & E. T. Luthfi, *Algoritma Data Mining*. Yogyakarta: Andi, 2009.

[4] E. Pramudiono, "Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data," 2003. .

[5] A. G. Mabrur, "Penerapan Data Mining di Bidang Marketing untuk Memprediksi Potensi Kriteria Nasabah Menggunakan Metode Decision Tree di PD BPR Kabupaten Bandung Cabang Batujajar," Universitas Komputer Indonesia, 2011.

[6] S. & H. P. Kusumadewi, *Aplikasi Logika Fuzzy untuk Pendukung Keputusan*. Yogyakarta: Graha Ilmu, 2010.

[7] E. Irdiansyah, "Penerapan Data Mining pada Penjualan Produk Minuman di PT. Pepsi Cola Indobeverages Menggunakan Metode Clustering," Universitas Komputer Indonesia, 2010.

[8] A. Budiarti, "Aplikasi dan Analisis Data Mining pada Data Akademik," Universitas Indonesia, 2006.

[9] R. H. Tamba, "Penerapan Data MIning Menggunakan Algoritma Agglomerative Hierarchical Clustering untuk Segmentasi Data," Universitas Gadjah Mada, 2009.