Application of fuzzy C-means algorithm for basic school clustering in Samarinda city based on minimum educational service standard indicators

Karyo Budi Utomo¹, Arbain P², and Suminto³

¹Information and Technologi Enginering, Politeknik Negeri Samarinda, Indonesia
²Electrical Enginering, Politeknik Negeri Samarinda, Indonesia
³Busines Administration, Politeknik Negeri Samarinda, Indonesia

Email: kbu@poln.ac.id

Abstract. Quality of service in schools determines the level of success of education service. The minimum service standard or better known as (SPM) in education is one of the references for evaluating the performance of education services. There are six categories in SPM of education, namely facilities and infrastructure, educators and education staff, curriculum, education quality assurance, education assessment, and school management. Each category has an achievement indicator to be used as a reference to assess the achievement that has been done by each school. School grouping based on the Minimum Service Standards (SPM) of Education is expected to be a reference for the government in making policies and references for Education organizers to know their position and then take strategic steps to enter the best groups. One of these groupings is achieved using the Fuzzy C-Means clustering method which is a non-hierarchical Clustering algorithm. The results of the study are expected to obtain a model that stakeholders can use in conducting elementary school groupings based on minimum service standard (SPM) in Education which are divided into 3 (three) categories, namely: Less, Medium, and Good.

1. Introduction
The 1945 Constitution Article 31 paragraph (1) mandates that every citizen has the right to education. The 1945 Constitution Article 31 paragraph (2), every citizen is required to attend basic education and the government is obliged to finance it [1]. To improve the quality of adequate education, a Minimum Service Standard (SPM) is needed. Regulation of the Minister of Education and Culture of the Republic of Indonesia, Number 23 of 2013 Article 6 paragraph (1), Education Minimum Service Standards (SPM) education, is a reference in program planning and budgeting achievement of targets in each regency/city [2].

Data from Samarinda City Education Office in 2015 showed that there were 184 public primary schools in Samarinda. The government has declared that every school must meet the SPM. There are 6 (six) categories of service types that become the minimum standard reference, namely: 1) Facilities and Infrastructure, 2) Educators and Education Personnel, 3) Curriculum, 4) Educational Quality Assurance, [1] Educational Assessment, and 6) School Management [2]. Each category has an achievement indicator to be used as a reference to assess the achievement of MSS that have been carried out by each school.

A large amount of data can slow down the clustering process of primary schools with good minimum service standards. The clustering of public elementary schools will be divided into three clusters, namely: poor, medium, and good.

The method used is fuzzy c-means, fuzzy c-means is a data clustering technique where the existence of each data point of a cluster is determined by the value of membership. The membership value will include real numbers at intervals of 0-1 [3,4]. The advantage of the fuzzy c-means method is that it is
capable of grouping irregularly distributed data [5,6]. The cluster formed is influenced by the input of several inputs in the fuzzy c-means process, such as the number of iterations, ranks, and the smallest error, but is not affected by the objective function and the initial iteration [5].

2. Review of Theory

2.1. Minimal Service Standards (SPM)
According to the Regulation of the Minister of Education and Culture of the Republic of Indonesia Number 23 of 2013, Article 6 [2]:
(1) SPM education is a reference in program planning and budgeting the achievement of targets in each district/city. (2) Program planning and budgeting for education Minimum service Standard as referred to in paragraph (1) shall be carried out in accordance with established technical guidelines/standards. (3) The target of achieving basic services in education must be achieved by the end of 2014. SPM education has six categories of service types, as described above.

2.2. Data Mining
Data mining is a method of processing data to find hidden patterns from the data. The results of data processing with this data mining method can be used to make decisions in the future. Data mining is data processing on a large scale, so data mining has an important role in the fields of industry, finance, weather, science and technology [7]. Data mining can also be done in various types of databases and information storage, but the type of pattern that will be found is determined by various data mining functions such as class/concept description, association, correlation analysis, classification, prediction, cluster analysis and others. Data mining uses the application of certain algorithms to extract patterns from data. Where this process will automatically look for simple patterns of large data using a particular analysis. Data mining also uses sophisticated mathematical algorithms to segment data and evaluate the possibility of some results set by the user [8].

2.3. Clustering
Clustering is the process of grouping objects based on information obtained from data that explains the relationships between objects with the principle of maximizing the similarity between members of one class and minimizing similarities between classes or clusters [9]. There are two stages that must be done in cluster analysis, namely: 1) Deciding whether the number of clusters is determined or not, and 2) Determine the algorithm to be used in clustering. To decide how many clusters to be formed [10]. In conducting clustering analysis can choose one of two approaches, namely 1) Hard Clustering or 2) Soft Clustering. The choice of the approach used depends on the type of data to be grouped. Hard clustering is used if the data is crisp, while soft clustering is used when the data is fuzzy [11].

2.4. Fuzzy C-Means
The Fuzzy C-Means algorithm was first proposed by Dunn in 1973 and then updated by Bezdek in 1981. This algorithm is one of the most popular soft clustering techniques using a data point approach where the cluster center point will always be updated according to the membership value of existing data and besides the fuzzy c-means algorithm is also an algorithm that works using the fuzzy model so that it allows all data from all group members to form with different degrees of membership between 0 and 1 [12][13]. Following is the elaboration of the FCM algorithm [14]:

Enter data to be clustered in the form of a matrix X size n x m (n = number of sample data and m = number of variables per data). Xij = i-th sample data (i = 1, 2, ..., n), the j-th variable (j = 1, 2, ..., m).

- Determine:
  - Number of clusters to be formed (c)
  - The weighting device (w)
  - Maximum iteration (MaxItr).
  - The smallest error
  - The initial objective function (P0 = 0)
  - Initial iteration (t = 1)

- Generating the initial partition matrix Unxc = [µik], µik is a random number representing a degree of membership

- Calculate the center of the k-th cluster (Vkj) with k = 1, 2, ..., c; and j = 1, 2, ..., m as follows:
  \[ V_{kj} = \frac{\sum_{i=1}^{n} (\mu_{ik})^w X_{ij}}{\sum_{i=1}^{n} (\mu_{ik})^w} \]

- Calculate the objective function on the t-iteration, Pt, which describes the amount of data distance to the center of the cluster.
  \[ V_k = \sum_{i=1}^{n} \sum_{k=1}^{c} \left( \frac{\sum_{j=1}^{m} (X_{ij} - V_{kj})^2}{\mu_{ik}^w} \right) \]

- Fix the degree of partition matrix membership:
  \[ \mu_{ik} = \frac{\left[ \sum_{j=1}^{m} (X_{ij} - V_{kj})^2 \right]^{-1}}{\sum_{i=1}^{n} \left[ \sum_{j=1}^{m} (X_{ij} - V_{kj})^2 \right]^{-1}} \]

- Check the stop condition
  - If you know (t > MaxIter) then stop
  - If not: t = t + 1, repeat step 4

2.5. Cluster Validity Testing

The results of clustering were tested for their level of validity by using the Modified Partition Coefficient measurement method to measure the level of cluster validity using values of µic membership degrees [15]. The main purpose of cluster validity is to evaluate the quality of the cluster and determine how well the data is represented by the selected cluster [16]. MPC is an improvement of the Partition Coefficient (PC) method. The PC method proposed by Bezdek in [17] is defined by Equation (4).
\[
PC(c) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})^2
\]

(4)

Where:
- \( c \) = number of clusters
- \( N \) = amount of data
- \( \mu_{ij} \) = degree of \( k \)-\( j \) data membership in the \( i \) cluster
- \( PC(c) \) = \( PC \) index value in the \( c \)th cluster

The partition coefficient tends to experience a monotonous change in various values of \( c \) (number of clusters) [18]. Modifications from the \( PC \) (Modified Partition Coefficient / MPC) index can reduce these monotonous changes. The MPC value is within the limit of \( 0 \leq PC(c) \leq 1 \). Generally, the most optimal number of clusters is determined from the highest MPC value \((\max_2 \max_{c\leq n-1} PC(c))\). Following is the MPC method algorithm: [18].

\[
MPC(c) = 1 - \frac{c}{c-1} (1 - PC(c))
\]

(5)

Where:
- \( c \) = number of clusters
- \( MPC(c) \) = MPC index value in the \( c \)-cluster

3. Research Methods

The research methods that will be carried out can be grouped into several sections, including:

3.1. Initial research

Studying matters related to the research topic. The main parts that need to be studied and explored include minimum education service standards (SPM), the basics of clustering, as well as an understanding of the Fuzzy C-Means method.

3.2. Data Collection

The data collected for this study is an indicator of the achievement of SPM (Minimum Service Standards) at Elementary Schools sourced from the Samarinda City Education and Culture Office.

3.3. Data analysis

Data collected, then analyzed based on research needs, among others, by observing the data structure and looking for patterns of data adjustment based on research needs. In each minimum service standard (SPM) category, it has several achievement indicators; the clustering process is carried out based on the SPM category so that the values on each achievement indicator are added up to then become a category value.

3.4. Application of clustering

At this stage, the clustering process is carried out using the fuzzy c-means method, the steps in the clustering process as previously described.

3.5. Evaluation

At this stage, the results of clustering are tested for validity using the Modified Partition Coefficient measurement method to determine the optimal number of clusters. Cluster segmentation results obtained from the results of the Fuzzy C-means clustering algorithm are compared with other period datasets to measure the performance of the mining process.

4. Results and Discussion

As explained earlier, the number of public primary schools to be clustered in 184 schools and the number of clusters is 3. The test in this study was carried out with a number of scenarios that were carried out by changing the input on several inputs in the fuzzy c-means process. The input that is changed is the maximum iteration, rank, and smallest error. All inputs are carried out with the same number of clusters,
the number of clusters is an assessment of the results of grouping, which in this study uses three clusters, namely: good, moderate and less.

4.1. Testing based on maximum iteration

Testing the maximum iteration is done with the aim to determine the effect of iteration on the results of the cluster. In this test, the number of clusters is 3, with rank weights is two and the smallest error is 10^-7 (0.0000001), and the initial objective function is 0. The testing process is carried out 10 times, by taking random iterations which include iteration 1; 5; 10; 15; 20; 30; 40; 45; 50; and iteration 58.

| Iteration number | Result of Cluster |
|------------------|-------------------|
|                  | Good  | Moderate | Less  |
| 1                | 38    | 77       | 69    |
| 5                | 47    | 69       | 68    |
| 10               | 68    | 61       | 55    |
| 15               | 75    | 64       | 45    |
| 20               | 75    | 67       | 42    |
| 30               | 73    | 66       | 45    |
| 40               | 73    | 66       | 45    |
| 45               | 73    | 66       | 45    |
| 50               | 73    | 66       | 45    |
| 58               | 73    | 66       | 45    |

Table 1 shows that from 10 scenarios, the maximum iteration data collection, at the maximum iteration 1; 5; 10; 15; and 20 shows the data distribution of cluster membership is still not definitive or still changing with each other. At a maximum of 30 iterations; 40; 45; 50; and 58 indicate that the cluster membership distribution level is definitive, and there is no change to the smallest error limit of 10^-7 on the 58th iteration.

4.2. Testing based on rank

Testing of rank values is carried out in order to determine the effect of rank on cluster results. In this test, the number of clusters is 3, with a maximum iteration is 100, the smallest error is 10^-7 (0.0000001), the initial objective function is 0 with the initial iteration 1. The testing process is carried out 10 times, with rank values including, 1; 2; 3; 4; 5; 6; 7; 8; 9; 10.

| Instrument | Result of Cluster |
|------------|-------------------|
|            | Good  | Moderate | Less  |
| 1          | 0     | 109      | 75    |
| 2          | 73    | 66       | 45    |
| 3          | 75    | 68       | 41    |
| 4          | 77    | 78       | 29    |
| 5          | 9     | 126      | 49    |
| 6          | 27    | 79       | 78    |
| 7          | 37    | 97       | 50    |
| 8          | 37    | 99       | 48    |
| 9          | 37    | 100      | 47    |
| 10         | 56    | 66       | 62    |

Table 2. Comparison of test results with different ranks
Table 2 shows that of the 10 data collection scenarios based on rank, starting from rank 1 to rank 10, the value of the members in each cluster varies.

4.3. Testing Based on the Smallest Error

Testing the smallest error is done with the aim to determine the effect of the smallest error on the results of the cluster. In this test the number of clusters is 3, with a maximum iteration is 100, the initial objective function is 0 and with an initial iteration one and with a rank value 2. The testing process is carried out seven times, including the smallest error 0.1; 0.01; 0.001; 0.0001; 0.00001; 0.000001; and 0.0000001.

Table 3. Comparison of test results with the smallest of different errors

| The smallest error | Good | Moderate | Less |
|--------------------|------|----------|------|
| 0.1                | 75   | 67       | 42   |
| 0.01               | 73   | 66       | 45   |
| 0.001              | 73   | 66       | 45   |
| 0.0001             | 73   | 66       | 45   |
| 0.00001            | 73   | 66       | 45   |
| 0.000001           | 73   | 66       | 45   |
| 0.0000001          | 73   | 66       | 45   |

Table 3 shows the seven scenarios for taking the smallest error data, the smallest error value of 0.1 occurs in the 20th iteration, then in the smallest error condition 0.01 until the smallest error, 0.0000001 condition distribution of member values in each cluster does not change.

4.4. Clustering Validity Testing

In testing the validity of clustering, the number of clusters is 3, with a maximum iteration of 100, the initial objective function is 0, with an initial iteration of 1 and with a rank of 2, and with the smallest error of 0.0000001. The test is carried out using the Partition Coefficient method and the Modified Partition Coefficient method. The testing mechanism is carried out with variations of 10% data (18 data), 25% (46 data), 50% (92 data), 75% (138), and 100% (184) of a total of 184 data. With random data collection.

Table 4. Testing Clustering Validity with Comparison of data variations

| Number of Data set | Method of Test | PC   | MPC  |
|--------------------|----------------|------|------|
| 100% (184)         | 0.4797         | 0.2195 |
| 75% (138)          | 0.5152         | 0.2729 |
| 50% (92)           | 0.5252         | 0.2879 |
| 25% (46)           | 0.5303         | 0.2955 |
| 10% (18)           | 0.5764         | 0.3646 |

Table 4 shows with all the scenarios the amount of data variation, in the Partition Coefficient method, the validity of the clustering has a tendency of optimal validity. All data variations worth more than 50% are close to 1. While the Modified Partition Coefficient method of clustering validity has a tendency of suboptimal validity because all data variations are worth less than 50% close to 1 (one), meaning that all data variation values are closer to 0. In this test, both using the Partition Coefficient method and the Modified Partition Coefficient method show that with a smaller number of data sets, the potential for validity is optimal.

4.5. Analysis of Testing Results
Based on testing to determine the performance of the fuzzy c-means algorithm for clustering the following results are obtained:

- Input in some inputs in the fuzzy c-means process affects the output of the cluster. The scanned input to be changed is the maximum iteration, rank and smallest error. The maximum value of iteration and smallest error with different input scenarios affects the members of each cluster but still produces the number of clusters as planned. Whereas in rank testing scenarios in addition to affecting the number of members of each cluster also affect the number of clusters, it happens when the rank value is 1.

- Testing based on the maximum iteration, the results vary. The maximum iteration 1 was 38.3 good schools, 77 schools and 69 schools were lacking. The number of iterations is too small, so the results are not right because the center of the cluster has not moved to the right position. Likewise, iterations in 5, 10 and 20 still obtain results that are varied in each cluster, in the iteration not yet obtained maximum results because the iteration is too little and does not meet the conditions according to the rules of the fuzzy c-means algorithm (t> MaxIter). Whereas in the 30, 40, 45, 50 and 58 iterations, the number of clusters and the number of members of each cluster is the same, namely the good category of 73 schools, 66 schools, and 45 schools less; this is because the iteration will stop if the objective function value during iteration to t and the objective function during the previous iteration (|Pt - Pt-1 | <ξ) the difference is smaller with the smallest expected error. So the exact center position of the cluster center is between the 20th and 30th iterations so that in more than 30 iterations, the same results will be obtained.

- Testing is based on the number of ranks, from 10 test scenarios, on testing with rank 1 produces 2 clusters even though it is conditioned by using 3 clusters. Whereas the testing with rank values 2 to 10 produces 3 clusters with different members of each cluster.

- At the test, based on the smallest error, the results vary. Only the smallest error value of 0.1 has a different number of members in each cluster, which is a good category of 75 schools, 67 schools and 42 schools less. While the smallest error is 0.01 to 0.0000001, the number of members in each cluster has similarities; that is, 73 schools are good, 66 schools are moderate and 45 schools are less.

- In testing the validity of clustering with 3 cluster scenarios tested based on the number of varied data sets divided into five variations of the amount of data, namely 10%, 25%, 50%, 75% and 100% of the total data of 184 schools, and conducted by using the Partition Coefficient and Modified Partition Coefficient methods. Testing uses the Partition Coefficient method: the validity of the clustering has a tendency for optimal validity. All data variations worth more than 50% are close to 1. While the Modified Partition Coefficient method of clustering validity has a tendency of suboptimal validity because all data variations are worth less than 50% close to 1 (one), meaning that all data variation values are closer to 0. In this test, both using the Partition Coefficient method and the Modified Partition Coefficient method show that with a smaller amount of data set, the potential validity will be increasingly optimal.

5. Conclusions
Based on the results of the discussion, the following conclusions are obtained:

- The fuzzy c-means method can be applied in clusters, variations in the input of several inputs in addition to affecting the value of each cluster member also affect the number of clusters formed

- The rank value affects the number of clusters, even though the number of clusters has been determined as many as 3 clusters, in the rank value 1, there are 2 clusters, which are medium and low clusters.

- The more maximum iterations and the smaller the error value of the cluster center will be in the right position

- Test the validity of the Partition Coefficient method, cluster validity has a tendency for optimal validity, whereas the Modified Partition Coefficient method has cluster validity that tends to be less than optimal.

- The less amount of data, both using the Partition Coefficient method and the Modified Partition Coefficient method, the potential for validity will be more optima
Reference

[1] Anon 1945 The 1945 Constitution of the Republic of Indonesia (Republic of Indonesia)
[2] Kemdikbud 2013 [2]. Regulation of the Minister of Education and Culture Number 23 of 2013 concerning Amendments to the Regulation of the Minister of National Education Number 15 of 2010 Minimum Service Standards for Basic Education in Regencies / Cities
[3] Megawati N, Mukid M A and Rahmawati R 2013 Segmentasi Pasar Pada Pusat Perbelanjaan Menggunakan Fuzzy C-Means (Studi Kasus: Ria Pasaraya Cilacap) J. Gaussian 2 343–50
[4] Priandini N, Zaman B and Purwanti E 2017 Categorizing document by fuzzy C-Means and K-nearest neighbors approach AIP Conference Proceedings vol 1867 (AIP Publishing LLC) p 20012
[5] Ahmadi A and Hartati S 2013 Application of Fuzzy C-Means in Decision Support Systems for the Determination of PNPM-MPd Community Direct Assistance Recipients (BLM) (Case Study of PNPM-MPd Ngadirojo District Pacitan) mencanangkan Program Nasional Pemberdayaan Mas Berk. MIPA 23 264–73
[6] Kustiyahningsih Y and Suprajitno H 2017 Development of an interval value fuzzy number based on MCGDM model by hybrid AHP and TOPSIS methods 2017 International Conference on Sustainable Information Engineering and Technology (SIET) (IEEE) pp 62–7
[7] Ong J O 2013 Implementasi algoritma k-means clustering untuk menentukan strategi marketing president university
[8] Ndehedhe C, Ogunlade S and Akwaowo E 2013 Spatial image data mining using K-Means analysis: a case study of Uyo Capital City, Nigeria Int. J. Adv. Res. 1 6–15
[9] Agusta Y 2016 Clustering
[10] Kusumadewi S and Purnomo H 2010 Aplikasi Logika Fuzzy untuk pendukung keputusan Yogyakarta Graha Ilmu 33–4
[11] Gan G, Ma C and Wu J 2007 Data clustering: theory, algorithms, and applications vol 20 (Siam)
[12] Bora D J, Gupta D and Kumar A 2014 A comparative study between fuzzy clustering algorithm and hard clustering algorithm arXiv Prepr. arXiv1404.6059
[13] Sanmorino A 2012 Clustering batik images using fuzzy c-means algorithm based on log-average luminance Comput. Eng. Appl. J. 1 25–31
[14] Andriyani T., Linawati L and Setiawan A 2013 Application of Fuzzy C-Means (FCM) Algorithm in Determining the Location of Establishment of the Salatiga PDAM Water Payment Counter in Proceedings of the SWCU Sci. Educ. Sci. VIII Salatiga
[15] Amirah M M A, Widodo A W and Dewi C 2017 Pengelompokan Lagu Berdasarkan Emosi Menggunakan Algoritma Fuzzy C-Means J. Pengemb. Teknol. Inf. dan Ilmu Komput. e-ISSN 2548 964X
[16] Sandhir R P, Muhuri S and Nayak T K 2012 Dynamic fuzzy c-means (dFCM) clustering and its application to calorimetric data reconstruction in high-energy physics Nucl. instruments methods Phys. Res. Sect. A Accel. spectrometers, Detect. Assoc. Equip. 681 34–43
[17] Maimon O and Rokach L 2005 Data mining and knowledge discovery handbook
[18] Nugraheni Y 2013 Data Mining Using Fuzzy Method for Customer Relationship Management in Retail Industry Lontar Komput. J. Ilm. Teknol. Inf.