Student academic performance analysis using fuzzy C-means clustering

R. Rosadi¹, Akamal¹, R. Sudrajat¹, B. Kharismawan², Y.A. Hambali³

¹Department of Computer Science, Padjadjaran University, Sumedang 45363, Indonesia
²Department of Mathematics, Padjadjaran University, Sumedang 45363, Indonesia
³Master Student of STEI, Institute of Technology Bandung, Bandung 40132, Indonesia

E-mail: r.rosadi@unpad.ac.id, akmal@unpad.ac.id, r.sudrajat@unpad.ac.id, bagus.kharismawan@unpad.ac.id, yudiahmadh@gmail.com

Abstract. Grade Point Average (GPA) is commonly used as an indicator of academic performance. Academic performance evaluations is a basic way to evaluate the progression of student performance, when evaluating student’s academic performance, there are occasion where the student data is grouped especially when the amounts of data is large. Thus, the pattern of data relationship within and among groups can be revealed. Grouping data can be done by using clustering method, where one of the methods is the Fuzzy C-Means algorithm. Furthermore, this algorithm is then applied to a set of student data form the Faculty of Mathematics and Natural Sciences, Padjadjaran University.

1. Introduction
Evaluation of academic performance is one of the bases for monitoring the development of students' academic achievement in higher education level. In evaluating the student academic performance data, there are occasion where the student data is grouped especially when it involves a large amounts of data. Thus, the pattern of data relationship within and among groups can be revealed. With the help of clustering techniques, student data can be classified based on their academic achievement. Clustering is one method of data mining where the process of grouping data objects into different classes called clusters so that objects that are in a cluster have differences with objects in other clusters. One of the methods used in clustering is the so-called fuzzy C-means algorithm. This algorithm can be used to collect a set of objects into a desired cluster. The authors have done before researches relating to the academic data in 2015. It was classified the academic data through several methods, such as, K-Means, KNN, FCM, and others. A web-based mobile application has also been developed as a tool for information processing and presentation of results.

2. Method
2.1 Data Mining
Data Mining refers to a process or method for extracting or mining knowledge or patterns from large amounts of data. Data mining emerged due to the availability of large amounts of data and the need to transform the data into useful information or knowledge [1]. To perform the data mining process, the concept of KDD (Knowledge Discovery in Databases) is commonly needed, see Figure 1.
2.2 Fuzzy Set
Fuzzy logic was first introduced by Lotfi A. Zadeh in 1965 (Zimmerman, 1991). Zadeh’s definition of fuzzy sets is as follows: If $X$ is a collection of objects denoted generically by $x$, then a fuzzy set $A$ in $X$ is defined as a set of ordered pairs:

$$A = \{(x, \mu_A(x)) \mid x \in X\},$$

with $\mu_A(x)$ is called the membership function for the fuzzy set $A$. The membership function maps each element of $X$ to a membership grade between 0 and 1 [2, 4].

2.3 Fuzzy C-Means Clustering
Fuzzy C-Means is an algorithm of data clustering, where each data becomes a member of a cluster with a degree defined by membership function [2]. Clustering process using Fuzzy C-Means algorithm has the following steps:
1. The data is stored $X$, a matrix of size $n \times m$, where $n$ is the number of samples and $m$ is the number of attributes.
2. Determine the number of clusters ($c$), rank ($W$), the maximum iteration (MaxIter), smallest error expected ($\varepsilon$), the initial objective function ($P_0 = 0$), and the initial iteration ($t = 1$).
3. Generate a random number $\mu_{ik}$ for $i = 1, 2, ..., n$ and $k = 1, 2, ..., c$ as an element of the matrix initial partition $U$. For $i = 1, 2, ..., n$, calculate

$$Q_i = \sum_{k=1}^{n} (\mu_{ik})^w$$

$$\mu_{ik} = \frac{\mu_{ik}}{Q_i}$$

4. Calculate the cluster center $V_{kj}$, with $k = 1, 2, ..., c$ and $j = 1, 2, ..., m$

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

5. Calculate the objective function at iteration $t$

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

6. Calculate the partition matrix transition
\[ \mu_{ik} = \frac{\left[ \sum_{j=1}^{m} (X_{ij} - V_{kij})^2 \right]^{-1}}{\sum_{k=1}^{c} \left[ \sum_{j=1}^{m} (X_{ij} - V_{kij})^2 \right]^{-1}} \] (6)

7. If \(|P_t - P_{t-1}| < \varepsilon \) or \(t > \text{MaxIter} \) then stop, otherwise \(t = t + 1\), repeat step 4.

Steps in Fuzzy C-Means algorithm can be illustrated by the following diagram (Figure 2).

2.4 Student Academic Performance
Academic performance is a result achieved in the academic field. Evaluation of students' academic performance can be seen from the GPA (grade point average) student. GPA is a number that shows achievement or academic performance of students cumulatively starting from the first semester [5].

3. Result and Discussion
Implementation of Fuzzy C-Means method can be seen in the trial grouping data using a web-based mobile application. In this study, the amount of data used is as many as 4255 records which to be grouped into four clusters, \(w = 2\), and the minimum error is 0.0001. Once the data to be grouped has been stored into the database, then the clustering process can be done. The process and the results of clustering that
has been done by the web-based mobile applications is provided in Figure 3 and the 4-clusters can be seen in Table 1.

![Image](image-url)

Figure 3. Mobile Web Application Process of Fuzzy C-Means Clustering

| Cluster | Centroid of GPA | Centroid of Study duration | Category |
|---------|-----------------|----------------------------|----------|
| 1       | 2.85            | 7.05                       | Bad      |
| 2       | 2.99            | 5.52                       | Not Good |
| 3       | 3.33            | 3.85                       | Very Good|
| 4       | 3.13            | 4.57                       | Good     |

The member of each cluster can be presented by the application in the form of graphs, which is shown in Figure 4. Figures 5-9 show the clusters with respect to the origin of department.

4. Conclusion

Fuzzy C-Means method can be implemented into the program to classify student academic data. Based on the results of the clustering process of 4255 student academic data using applications. Four clusters suitable for use as a result of grouping data for each cluster of students do not have adjacent centroid that can have different categories. Cluster with the best academic performance categories, namely Cluster 3 totaling 1753 students, followed by a Cluster 4 totaling 1496 students, Cluster 2 totaling 676 students, and Cluster 1 totaling 330 students. From the results of the grouping that is associated with a department, the percentage of students that included in Cluster 1 (unfavorable category), the highest is the Department of Mathematics, Cluster 2 (category quite well) were the highest is Department of Biology, Cluster 3 (category Excellent) that the highest is Department of Statistics, Cluster 4 (both categories) were the highest is Department of Biology. Especially for the Department of Information Engineering and Geophysics, the condition of the current data that is processed at this time fewer new
graduates, S1-Informatics 3 graduates, where all three go in Cluster 3 (excellent category) and S1 Engineering Geophysics 40 graduates, of which 31 graduates entered Cluster 3 (excellent category) and 9 graduates enter the Cluster 4 (both categories).

Figure 4. The resulting graph FCM Clustering

Figure 5. Cluster 1 based on department

Figure 6. Cluster 2 based on department
Figure 7. Cluster 3 based on department

Figure 8. Cluster 4 based on department

Acknowledgment
Authors thank the organizers of DRPM Unpad for the HPKRD research grants, and FMPA Unpad as a facilitator. Author thank also other parties who have contributed a lot in research and completion of the seminar paper.

References
[1] Han, J. Kamber, M. 2006. Data Mining Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.
[2] Jang, JSR Sun, CT dan Mizutani, E. 2004. Neuro Fuzzy and Soft Computing. Singapore. Pearson Education.
[3] Kharismawan, Bagus. 2015, Aplikasi K-Means dan Fuzzy C-Means Clustering untuk Mengelompokkan Data Mahasiswa FMIPA Unpad: University of Padjadjaran.
[4] Kusrini, Luthfi. 2009, Algoritma Data Mining. Yogyakarta: Andi Offset.
[5] http://www.unpad.ac.id/pengajaran/evaluasi-hasil-belajar-dan-batas-waktu-studi/evaluasi-hasil-belajar/