Application of fuzzy C-Means Algorithm for Determining Field of Interest in Information System Study STTH Medan

Edy Rahman Syahputra¹, Yulia Agustina Dalimunthe², Irvan³

¹,²Department of Information System
Sekolah Tinggi Teknik Harapan, Medan-20217, Indonesia
³Department of Mathematics Education
Universitas Muhammadiyah Sumatera Utara, Medan Indonesia

¹ydeaja@yahoo.com, ²yuliaagustina_dalimunthe@yahoo.co.id, ³irvan@umsu.ac.id

Abstract. Many students are confused in choosing their own field of specialization, ultimately choosing areas of specialization that are incompatible with a variety of reasons such as just following a friend or because of the area of interest of many choices without knowing whether they have Competencies in the chosen field of interest. This research aims to apply Clustering method with Fuzzy C-means algorithm to classify students in the chosen interest field. The Fuzzy C-Means algorithm is one of the easiest and often used algorithms in data grouping techniques because it makes efficient estimates and does not require many parameters. Several studies have led to the conclusion that the Fuzzy C-Means algorithm can be used to group data based on certain attributes. In this research will be used Fuzzy C-Means algorithm to classify student data based on the value of core subjects in the selection of specialization field. This study also tested the accuracy of the Fuzzy C-Means algorithm in the determination of interest area. The study was conducted on the STT-Harapan Medan Information System Study program, and the object of research is the value of all students of STT-Harapan Medan Information System Study Program 2012. From this research, it is expected to get the specialization field, according to the students' ability based on the prerequisite principal value.

1. Introduction

A change or transformation of a curriculum in the field of education continues to occur, it is intended to make a better change in the world of education in Indonesia, not least in the Information Systems Study Program at Sekolah Tinggi Teknik Harapan. In 2015 the Information Systems Study Program implements a new curriculum based on KKNI (Indonesia National Qualification Framework) which in the new curriculum requires students to choose an area of specialization that exists. There are 3 areas of specialization in the new curriculum are: (1) System Analysis, (2) Information Management, (3) Intelligent Information System. With the field of specialization that must be selected by students raises a problem of students, especially students that are in the Level III / student class of 2012. All grades that have been taken by students that are in level III have been converted into new curriculum so as to cause all the prerequisite courses of the field of specialization has been valuable, and because it is finally the Information System Study Program to make a policy by giving full rights to the students to Choose your own specialized field. From the results of the survey conducted by researchers in the field, many students that are confused must choose their own specialized area which, finally choosing a specialized area that is not appropriate with a variety of reasons such as just following a friend or
because the specialized field that many choose without knowing if the self is competent in the field of selection that has been selected. Therefore, it needs recommendations of them, to recommend so many students is not easy because we have to manage the data so large that has the number of fields and the number of records that so much. Data mining technology is one of the tools for extracting data onto large databases and with complexity specifications that have been widely used in many application domains such as banking and telecommunications. In data mining there is a technique in managing large data that is clustering technique. Clustering is a grouping with unlike objects of other clusters. The principle of clustering is to maximize the homogeneity/similarity between traits within a group and to maximize the heterogeneity/inequality of traits between groups. In addition, this method is able to group data scattered irregularly. Several studies conducted research on clustering data onto various methods used, among others, research conducted by some of the following researchers and research determination of high school majors at C-Means Algorithm have higher accuracy (that is average 78.39%), if compared with method of determination majors manually which have been done only have the level of accuracy of average 56.17. Determining the ideal cluster by using variables in research that discusses the implementation of the single linkage method to determine agent performance on call centres based on asterisks for Java [1][2]. Therefore, the cluster technique with the fuzzy c-means algorithm can be used to classify the students based on the prerequisite subject value related to the specialization field so that it can give an idea of the students to choose the program of interest according to them.

2. Literature Review

2.1. Clustering Concept in Data Mining

The basic concept of data mining is to find hidden information about a database and to be part of Knowledge Discovery in Database (KDD) to find useful information and patterns in the data [3][4]. Clustering divides data onto groups or clusters based on a similarity in attributes between the data [5][6]. Characteristic of each cluster is not Determined before, but reflected by the similarity of data grouped in it. Therefore, clustering results often needs to be interpreted by the parties that really understand about the character of the data domain. In addition to being used as an independent method of data mining, clustering is also used in pre-processing of data before data is processed by other data mining methods to improve the data domain. The most important characteristic of a good clustering result is that a data instance in a cluster is more "similar" to another instance within the clustering than with an instance outside of the clustering. Similarity measures can vary and affect the calculation in determining the member of a cluster. So the data type to be clustered (quantitative or qualitative) also to determine what size is appropriately used in an algorithm. In addition to the similarity between data onto a cluster, clustering can also be done based on the distance between data or cluster with one another. Distance size (distance or dissimilarity measure) which is the opposite of this similarity size is also a lot of variety and its usage also depends on the type of data to be clustered. Both sizes are symmetrical, where if A is said to be similar to B it can be concluded that B is similar to A. There are several kinds of distance calculation formulas for clusters. For the numerical data type, a data X of X, i = 1, ..., n, each item is represented as a vector X = \{X_i, X_{i2}, X_{im}\} with m as the number of dimensions of the item. The usual formulas used as a measure of the distance between X_i and X_j for this numerical data include.

2.2 C-Means Algorithm

In the classical clustering process (e.g. Clustering K-Means), the formation of partitions is done in such a way that each object is right on one partition, since the object is actually located between two or more other partitions. In algorithmic logic, a method that can be used to perform grouping of data is known as clustering algorithm. A clustering algorithm is said to be a clustering algorithm if the algorithm uses strategic parameters of adaptation soft competitive. Most clustering algorithms are based on the optimization of the objective function or modification of the objective function [13]. One of clustering algorithm technique is C-Means Algorithm. C-Means algorithm is a data clustering
technique that the existence of each data onto a cluster is determined by the value / degree of certain membership. This technique was first introduced by Jim Bezdek in 1981 [14]. In contrast to the classical clustering technique (where an object will only be a member of several clusters. The cluster boundaries of the C-Means algorithm are soft (soft). The basic concept of C-Means algorithm, the first is to determine the cluster center that marks the flat location for each cluster of the initial conditions, the center of the cluster is still not accurate. Each data has degree of membership of each cluster by repairing the cluster center and membership value of each data repeatedly, it will be seen that the center of the cluster will move toward the right location. This recurrence is based on the minimization of objective function. The objective function used for C-Means Algorithm are [15][16].

1. The C-Means algorithm is given as follows [17][18]: 1. Determining the data to be clustered X, in the form of nxm-sized matrix (n = number of data samples, m = attribute of each data), \( X_{ij} = \) i-sample data \( i = 1,2, ..., n \), attributes To-j (j = 1,2, ..., mm).

2. Determine:
   - Number of clusters = c
   - Rank = w - Maximum interaction = Maxlter
   - The smallest error expected = \( \xi \)
   - The initial objective function = \( P_0 = 0 \)
   - Initial interaction = t = 1

3. Generating random numbers \( \mu_{ik} \) i = 1,2,3, ..., n; k = 1,2,3, ...; c: as elements of the initial partition matrix U. Counting the number of each column: \( Q_l = \sum_k \mu_{ik} \)
   with \( j=1,2,..,n \)

4. Calculate the centre of the k-cluster: \( V_{kj} \), with k = 1,2, ..., c; and j = 1,2, m ...

5. Calculate the objective function of the interaction to-t:

6. Calculate the partition matrix changes:

7. By: i = 1,2, ..., n; and k = 1,2, ..., c. Checking the stop condition: - If: (| \( P_t - P_{t-1} | < \xi \)) or (t > Max)
   then stop - If not: t = t + 1, repeat step 4

3. Research Methodology

1. Identification of Problems: a preliminary study to examine the problems in the process of grouping students based on prerequisite courses in the selection process of specialization.

2. Approach in problem solving: is the stage of finding the right method in solving problems based on literature review. The Fuzzy C-Means Clustering algorithm that has been tested for several researches for the case of data grouping will be tested for classifying the preliminary subject value data onto the determination of the student interest area.

3. The tool used in modeling the sequence of activities in the using the activity diagram and for implementation system performed using software visual studio 2010. 

4. Implementation and testing for the system and results analysis: is the stage of the process of testing the results of the cluster fuzzy c-means method in classifying the field of student interest based on prerequisite course value, with data training in the form of prestigious student grade 2012.

5. Drawing the conclusions of the research results: after testing the system is completed to eat a result is drawn in the form of conclusions based on the results.

4. Implementation and Testing

The data used is the data value of students of the curriculum conversion results from 2011-2015 STT-Harapan Medan Information Systems Study Program as many as 11 classes. Parameter of data to be used in research in the form of the average of the subjects of the interest groups, namely: Specialization of Information Systems Analysis (Data Systems and Information Technology, Database, Programming Language, Business Process Analysis, Analysis and Design of Information Systems and Management of Technology Services Information, Information Management, Database,
Programming Languages, Probabilistics and Statistics, Object Oriented Programming and Business Modeling and Simulation System) and Intelligent Information System (Data Systems and Information Technology, Database, Programming Language, Operating System, Design and Management of Computer Network and Intelligent System). The sample data used in this research is Data onto Student Average Value of Special Interest Field Before Specialization and Data of student average value of Special Interest Field After Specialization.

**Measurement Method**

Accuracy of application of C-Means Algorithm in the specialization of its majors is the sample data onto the students and the average value of the subjects of interest before the interest grouped with the C-Means Algorithm to divide the students into the field of interest in accordance with the similarity in the average value of the field of interest. Then the outcome of the C-Means Algorithm is compared with the result of the interest in the sample data onto the average values of the subjects on the specialization that has been undertaken by the students. If the interest chosen by the student is equal to the concentration of C-Means Algorithm and the average value of the subjects acquired after specialization > = Minimum Criteria of Completeness (KKM) > (> = 75), the C-Means Algorithm is declared ACCURATE. If the interest chosen by the student is equal to the concentration of C-Means Algorithm and the average value of the specialization subject obtained after the <Minimum Exhaustiveness Criterion (> = 75), the C-Means Algorithm is NOT ACCURATE. If the interest chosen by the student is not the same as the C-Means Algorithm and the average course value obtained after the minimum > Minimum Criterion of Maximum Criterion, the C-Means Algorithm is NOT ACCURATE. If the interest obtained by the student is not the same as the C-Means Algorithm and the average value of the specialization subject obtained after the <Minimum Exhaustiveness Criterion (KKM) criterion, the C-Means Algorithm is declared ACCURATE.

Next calculated percent accuracy level Fuzzy C-Means with: \( \% \) Accuracy = (Accurate Total Data / Total Samples) * 100.

After the parameter of the average field value of interest is known, then mapping / clustering the data follows the FCM algorithm with setting the initial partitions matrix in the form of n x m matrix (n is the number of data samples, ie = 81, and m is the parameter / attribute of each data, ie = 3). \( X_{ij} \) is i-th sample data (i = 1,2, ..., n), j-attribute (j = 1,2, ..., m). Then determining Parameter Value Beginning. And followed by generating random numbers \( \mu_{ik} \), i = 1,2, ..., n; K = 1,2, ..., c; As the initial partition matrix element (U) formed by using Microsoft Visual Studio.

```matlab
>>rand('state',0)
>>X=[rand(7,3);rand(7,3);rand(7,3)]
```

Next step is define Cluster Centre (V) In the first iteration, using the equation:

\[
V = \begin{pmatrix} 
7.25 & 7.66 & 5.14 \\
7.88 & 7.79 & 6.53 \\
8.15 & 8.08 & 6.84 \\
\end{pmatrix}
\]
And then calculating the Objective Function (P)

\[
\begin{array}{cccc}
9.056941 & 0.005599 & 0.62408 & 9.68562
9.112247 & 43.289464 & 100.0489 & 153.0939
6.090798 & 70.9989 & 18.78422 & 201.8791
9.877825 & 237.2356 & 175.7346 & 402.8858
28.599 & 15.33861 & 67.87362 & 111.7772
94.85771 & 56.50128 & 0.099954 & 151.4589
15.47084 & 24.80597 & 55.06622 & 95.34503
137.2072 & 10.64514 & 42.35627 & 190.3087
45.0474 & 35.58129 & 9.122254 & 86.15094
38.67844 & 10.04519 & 138.6388 & 247.3525
13.89192 & 203.2413 & 190.9962 & 533.1567
176.7242 & 120.2178 & 188.2951 & 487.2011
25.91444 & 70.38042 & 0.432393 & 97.72783
65.87224 & 25.85041 & 91.81205 & 181.5747
13.84711 & 8.258158 & 19.48254 & 41.58825
44.36476 & 25.86134 & 21.73436 & 91.95046
127.2471 & 7.022698 & 238.109 & 273.2788
36.952 & 4.574086 & 162.6364 & 204.1625
0.0266056 & 0.06376 & 75.13057 & 75.23202
1050.238 & 1098.154 & 1567.618 & 5696.005
5.002971 & 144.6212 & 10.64904 & 160.2681
\end{array}
\]

The value of the objective function in the first iteration \((P_1)\) has a value of: Objective Function = 3696.005. 1. Checking Stop Conditions Because \(|P_1-P_0| = |3696.005-0| = 3696.005 >> \xi (10^{-5})\), and iteration = 1 < MaxIter (= 100), then proceed to second iteration \((t = 2)\). 2. Calculate the partition matrix changes with: \(i = 1, 2, \ldots, n\); and \(k = 1, 2, \ldots, c\)

Checking the stop condition:
At: \((|P_t-P_{t-1}| < \xi)\) or \((t > \text{Max})\) then stop in the 48th iteration with the cluster centre value:

\[
V = \begin{pmatrix}
7.57 & 7.55 & 7.49 \\
7.45 & 7.54 & 7.46 \\
7.56 & 7.53 & 7.54
\end{pmatrix}
\]

5. Conclusion
Clustering Method of Fuzzy C-Means Algorithm can classify students in the field of specialization according to the prerequisite course by utilizing UI output as determination of each student into which groups and utilize the output of Vi to determine each group is identified as specification. The Fuzzy C-Means algorithm uses repetitive iterations to specify the grouping. The aeration is repeated to the nearest optimum objective function, obtained at point \((|P_t-P_{t-1}| < \xi)\) or \((t > \text{Max})\) Then it stops at 48 iterations. This research builds an application for the application of special c-means algorithm in information system studies program. Hopefully in the next research can be built an application for determinants of interest in all majors at STT-Harapan Medan. For clustering just focusing on areas of interest expected in the future can be developed for the determinant of the student majors at the final grades in school at the top level or also the value of the UN prospective students.

References
[1] Anjana Gosain, Sonika Dahiya, Performance Analysis of Various Fuzzy Clustering Algorithms: A Review, Procedia Computer Science, Volume 79, 2016
[2] Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan et al. "Top 10 algorithms in data mining." Knowledge and Information Systems 14, no. 1 (2008).
[3] Kaur, Prabhjot, and Anjana Gosain. "Improving the performance of fuzzy clustering algorithms through outlier identification." In Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on, pp. 373-378. IEEE, 2009.
[4] Baoju Zhang, Shan Qin, Wei Wang, Dan Wang, Lei Xue, Data stream clustering based on Fuzzy C-Mean algorithm and entropy theory, Signal Processing, Volume 126, 2016
[5] Zhang BO, Fuzzy clustering based on the data flow model, Comput. Eng. Apl. (2010)
[6] Yinghua Shen, Witold Pedrycz, Collaborative fuzzy clustering algorithm: Some refinements, International Journal of Approximate Reasoning, Volume 86, 2017

[7] M. Zarinbal, M. H. Fazel Zarandi, I. B. Turksen, Relative entropy collaborative fuzzy clustering method, Pattern Recognit. 48 (2015) 933-940.

[8] Y. Jiang, F. Chung, S. Wang, Z. Deng, J. Wang, P. Qian, Collaborative fuzzy clustering from multiple weighted views, IEEE Trans. Cybern. 45 (2015)

[9] P. Rastin, G. Cabanes, N. Grozavu, Y. Bennani, Collaborative clustering: How to select the optimal collaborators? in: Proceedings of the 2015 IEEE Symposium Series on Computational Intelligence, 2016, pp. 787-794.

[10] C. T. Chu, H. K. Chiang and Y. S. Chang, "Clustering algorithms applied Gaussian basis function neural network compensator with fuzzy control for magnetic bearing system," 2016 5th International Symposium on Next-Generation Electronics (ISNE), Hsinchu, 2016, pp. 1-2.

[11] J. I. Sanchez-Gomez, L. Morales-Velazquez, R. A. Osornio-Rios and E. Guillen-Garcia, "Fuzzy c-means clustering for steady state events classification of electrical signals," 2016 12th Congreso Internacional de Ingeniería (CONIIN), Santiago de Queretaro, 2016, pp. 1-5.

[12] José Angel BARRIOS, César VILLANUEVA, Alberto CAVAZOS, Rafael COLAS, Fuzzy C-means Rule Generation for Fuzzy Entry Temperature Prediction in a Hot Strip Mill, Journal of Iron and Steel Research, International, Volume 23, Issue 2, 2016

[13] Engin Esme, Bekir Karlik, Fuzzy c-means based support vector machines classifier for perfume recognition, Applied Soft Computing, Volume 46, 2016

[14] Hanuman Verma, R.K. Agrawal, Aditi Sharan, An improved intuitionistic fuzzy c-means clustering algorithm incorporating local information for brain image segmentation, Applied Soft Computing, Volume 46, 2016

[15] Hassan Fathabadi, Power distribution network reconfiguration for power loss minimization using novel dynamic fuzzy c-means (dFCM) clustering based ANN approach, International Journal of Electrical Power & Energy Systems, Volume 78, 2016

[16] Jeetashree Aparajeeta, Pradipta Kumar Nanda, Niva Das, Modified possibilistic fuzzy C-means algorithms for segmentation of magnetic resonance image, Applied Soft Computing, Volume 41, 2016

[17] J. Aparajeeta, P.K. Nanda, N. Das, Bias field estimation and segmentation of MR image using modified fuzzy-C means algorithms, in: Proceedings of Eighth International Conference on Advances in Pattern Recognition (ICAPR), IEEE, 2015, pp. 1–6.

[18] Yi Ding, Xian Fu, Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm, Neurocomputing, Volume 188, 2016.