K-Means Algorithm to Group Students' Academic Status at STMIK Palangka Raya

Lili Rusdiana
Department of Informatics Engineering, STMIK Palangka Raya
e-mail: fasliiana7@gmail.com

Abstract

As one of clustering methods K-Means algorithm can be used to group 93 data of STMIK Palangka Raya students into 3 namely pass, active, and inactive. The stages of it produce 9 iterations since the displacement of the group in which there is a movement from one group to another because the change value of the object function. The iteration can be stopped at the 9th because the value of the object function change is below the given threshold value of 0.1 and the absence of group movement in the data used. The 9 iterations that occur indicate the rise and fall the change value of object function and data transfer in the location of the group.

***Kata kunci* :** *Grouping, K-Means Algorithm, Students' Academic Status*

1. INTRODUCTION

K-Means algorithm, as one of clustering methods, can be used to group a set of data based on certain characteristics contained in the set of the data. K-Means concept in classifying the data is that by dividing into two groups or more like dividing the quality of student learning into 6 categories through application built using K-Means algorithm to get the result as decision making to student learning progress so that its quality can be improved [1]. The previous research on application development to predict the predicate of students passing graduate using Fuzzy C-Means method concept to classify 2 clusters that pass and fail by using 98 data which randomly used as training data and test data. The result was that the more training data and test data the better the accuracy of the prediction using the Fuzzy C-Means method concept [2]. In addition, grouping the students' data was also done to determine the passing rate of 171 data used and displayed 30 samples of data with 3 clusters of length of study: on time, less timely, and not timely [3]. Students' data were again used to group performance into clever, standard, and low using criteria of GPA, gender, place of origin, course ever joined, and attendance [4]. The use of K-Means algorithm in determination of predicate of graduation of student was done but only using 10 random sample data to know result of calculation of K-Means algorithm and the result was 70% data could be recognized and 30% data couldn't [5]. Based on the previous research, this research is conducted to know the use of K-Means algorithm to the data set of Undergraduate program students of Information System Year 2011 with characteristic of GPA, number of credits that have been taken, and final project status that divided into 3 groups ie pass, active but has not passed, and inactive.

2. RESEARCH METHODOLOGY

2.1 Needs Analysis

Needs analysis requires several requirements regarding input data, process, and output. They are obtained from the data used to classify students based on their academic status at Information System study program of STMIK Palangka Raya 2011 that is got from academic administrator. Based on these data, the results of the analysis as follows:

1. Input Needs

Data used as inputs are students data such as students' number, name, credits, GPA, final project status, and graduation reference. But only the credits, GPA, final project status, and graduation reference that are processed using K-Means Algorithm.

2. Process Needs
   The process used to process the data input is K-Means algorithm and the implementation of using Microsoft Office Excel software.
3. Output Needs
   The expected output is the analysis of the clustering results that can recognize the information / pattern of the input data is that academic status such as inactive, active but not yet passed, and already passed because the data used is taken only to the limit of semester 10.

2.2 Grouping

One of the purposes of grouping is to find out a pattern of all data used. The data used in this research are 93 student data by applying K-Means algorithm and divided into 3 groups is that inactive, active but not yet pass, and have passed based on characteristic of GPA, number of credits, and status of final project.

Preliminary data obtained will be sorted into 3 groups based on the nearest characteristics or that have the sameness or similarities so that they can be separated by the data that do not. Some requirements to get the predicate of graduation achievement of Undergraduate program students based on Regulation of the Minister of Research, Technology and Higher Education of the Republic of Indonesia No. 44 of 2015 [6] then this research uses certain characteristics as follows:

1. Undergraduate students are stated pass if they have taken all the stipulated study load and have a graduate learning achievement targeted by the study program with acumulative grade point (GPA) greater than or equal to 2.00 (two point zero zero).
2. A maximum of 7 (seven) academic years for the undergraduate program, four diploma / applied undergraduate courses, with a student's learning load of at least 144 (one hundred and forty four) credits.
3. Research activities undertaken by students in order to carry out the final project or skripsi must meet the provisions of graduate learning achievements, and the provisions of regulations in universities.

There are some academic status of students of STMIK Palangka Raya such as pass, active, inactive, drop out, and others.

2.3 K-Means Algorithm

K-Means concept is one of the concepts in data mining that classifies by dividing data into two groups or more by aiming to minimize the object function that has been arranged in grouping process. K-Means algorithm in the data grouping as follows [7]:

1. Determine the number of groups.
2. Allocate data into groups at random.
3. Calculate the group center (average) of the data contained in each group.

The location of the center point of each group taken from the average of all data values on each feature must be recalculated. To calculate the i-feature centroid can be used as in equation 1.

$$C_i = \frac{1}{\mathrm{M}} \sum_{\mathrm{i}=1}^{\mathrm{M}} \mathrm{x_i} \qquad (1)$$

Where :
C = Centroid
M = Amount of data in a group
i = i-feature in a group
x = matrix data set M x N.
N = number of features
Equation 1 is done as much as p (dimensional data) dimension so *i* start 1 to p.

4. Allocate each data to the nearest centroid / average.
5. Back to step 3 if there is a datum moving group or there is a change of centroid value above the specified threshold value, or if the value change on the objective function used is still above the specified threshold value.

To measure the distance in the distance space by Euclidean is using the formula as in equation 2.

$$D(x_2, x_1) = ||x_2 - x_1||_2 = \sqrt{\sum_{i=1}^{p} |x_{2i} - x_{1i}|^2} \quad (2)$$

Where :
D = Distance between data x_2 AND x_1
| . | = Absolute value
p = Dimensional data
i = i-feature in a group
x = Matrix data set M x N.
N = Number of features

2.4 Implementation

Grouping using K-Means algorithm involves 93 data that has passed the preprosessing of 136 initial data to avoid double data. Grouping 93 data in 3 dimensional grouping data sets. The dimensional of 3 features are GPA, credits, and final project status. The value of each data is used as in table 1. The distance measurement used is Euclidean distance. Number of groups are 3: pass, active but not pass, and inactive. The threshold (T) used for the change of objective function is 0.1

Table 1. Value used for grouping

Number of Datum	GPA	Credits	Final Project
1	2	3	4
1	2,87	145	0,5
2	2,99	145	0,5
3	2,57	143	0,5
4	2,56	151	1
5	2,83	127	0,5
6	3,36	151	1
7	3,56	151	1
8	3,45	151	1
9	2,89	151	1
10	3,25	151	1
1	2	3	4
11	3,62	145	0,5
12	2,85	145	0,5
13	2,89	145	0,5
14	3,33	145	0,5
15	3,32	145	0,5
16	3,03	151	1
17	2,20	121	0,5
18	3,07	151	1

19	3,05	151	1
20	2,10	86	0,5
21	2,95	151	1
22	3,70	151	1
23	3,69	68	0,5
24	2,09	112	0,5
25	3,11	151	1
26	2,76	138	0,5
27	3,31	151	1
28	2,13	82	0,5
29	2,73	60	0,5
30	2,92	151	1
31	3,32	151	1
32	2,69	143	0,5
33	3,04	151	1
34	3,10	145	0,5
35	2,38	125	0,5
36	2,78	145	0,5
37	2,01	95	0,5
38	2,83	151	1
39	2,89	149	1
40	3,15	151	1
41	2,64	148	0,5
42	2,28	120	0,5
43	3,50	145	0,5
44	2,76	103	0,5
45	3,17	145	0,5
46	2,81	143	0,5
47	3,03	151	1
48	3,33	151	1
49	3,09	145	0,5
50	3,79	145	0,5
51	3,01	151	1
52	3,63	145	0,5
53	3,23	151	1
1	2	3	4
54	2,22	115	0,5
55	2,35	110	0,5
56	2,00	2	0
57	1,75	24	0
58	2,85	143	0,5
59	0,00	0	0
60	2,68	145	0,5
61	2,49	125	0,5
62	2,32	38	0
63	3,08	141	0,5

64	2,94	105	0,5
65	2,98	151	1
66	3,15	151	1
67	2,82	83	0,5
68	2,72	78	0,5
69	2,71	141	0,5
70	3,07	145	0,5
71	3,74	151	1
72	1,92	50	0,5
73	3,23	145	0,5
74	2,90	145	0,5
75	2,87	151	1
76	2,80	143	0,5
77	3,26	151	1
78	2,43	129	0,5
79	2,36	67	0,5
80	2,70	143	0,5
81	2,90	143	0,5
82	3,14	151	1
83	2,99	151	1
84	2,77	151	1
85	3,24	151	1
86	3,64	151	1
87	3,71	151	1
88	3,30	145	0,5
89	3,45	141	0,5
90	3,46	151	1
91	2,82	109	0
92	3,28	131	0,5
93	2,95	145	0,5

In table 1, the final project data are based on 3 sections is that 0 for not taking the final project, 1 for passed the final project, and 0.5 for doing the final project. The data used as in table 1 and carried out the following steps:

1. Initialization
   a. Allocate all data in one group at random
      From the randomization process that each data join to each group of 3 ie pass, active but not passed yet, and inactive.
   b. Count the group center (centroid)
      The central location of the group is the result from equation 1, to compute the group center of each feature taken from the average of all data joining in each group as in table 2 so produce the group center as in table 3.

Table 2. Counting group center of group 1

Number of Datum	GPA	Credits	Final Project
1	2	3	4
1	2,87	145	0,5

2	2,99	145	0,5
3	2,56	151	1
4	3,45	151	1
5	2,89	145	0,5
6	2,20	121	0,5
7	3,07	151	1
8	3,05	151	1
9	3,69	68	0,5
10	3,32	151	1
11	2,69	143	0,5
12	2,01	95	0,5
13	2,28	120	0,5
14	2,81	143	0,5
15	3,63	145	0,5
16	3,23	151	1
17	0,00	0	0
18	2,68	145	0,5
19	2,49	125	0,5
20	2,82	83	0,5
21	2,90	145	0,5
22	3,24	151	1
Average	2,77	128,41	0,64

Table 3. Group center of inisiation step

Group	GPA	Credits	Final Project
1	2,77	128,41	0,64
2	2,86	124,89	0,59
3	2,99	135,48	0,72

The average value in table 2 is used in table 3 to search center group and its result is to calculate the distance of data to the group center like equation 2 and got the objective function that equals to 2241,48 so that the change of objective function as follows:

Old distance = 0
New distance = 2241.48
Changes in objective function = | new distance - old distance |
= | 2241,48 - 0 |
= 2241.48

The value of object function change of 2241.48 is still above the threshold value of 0.1 then the center of the group is recalculated to the next iteration until the value of change of objective function is below the threshold or the absence of group movement.

2. Iteration 1
   a. Recalculate the shortest distance of each data in a particular group.
   b. Allocate any data on the nearest centroid
   c. Count the group center (centroid)

   As in the initialization process obtained group center for iteration 1 as in table 3 and the average value for group 1 as in table 4.

Table 4. Counting group center of group 1

Number of Datum	GPA	Credits	Final Project
1	2,83	127	0,5
2	2,43	129	0,5
3	3,28	131	0,5
Average	2,85	129,00	0,50

Table 5. Group center of Iteration 1

Group	GPA	Credits	Final Project
1	2,85	129,00	0,50
2	2,31	81,65	0,39
3	3,10	147,66	0,76

The average value in table 4 is used in table 5 and the objective function is 939,77 so that the objective function changes as follows:

Old distance = 2241,48
New distance = 939.77
Changes in objective function = | new distance - old distance |
= | 939.77 - 2241.48 |
= 1301,71

Go to the next iteration by directly showing the counting steps of group center.

3. Iteration 2
The process of iteration 2 results group center as in table 6.

Table 6. Group center of iteration 2

Group	GPA	Credits	Final Project
1	2,51	121,83	0,46
2	2,28	62,73	0,37
3	3,11	147,80	0,77

The objective function is equal to 656,49 so that the change of objective function as follows:

Old distance = 939.77
New distance = 656.49
Changes in objective function = | new distance - old distance |
= | 656,49 - 939,77 |
= 283.28

Go to the next iteration by directly showing the counting steps of group center.

4. Iteration 3
The process of iteration 3 results group center as in table 7.

Table 7. Group center of iteration 3

Group	GPA	Credits	Final Project
1	2,51	116,21	0,46
2	2,21	53,17	0,33
3	3,11	147,70	0,77

The objective function is 671,97 so that the objective function changes as follows:

Old distance = 656,49

New Distance = 671.97
Changes in objective function = | new distance - old distance |
= | 671,97 - 656,49 |
= 15.47

5. Iteration 4
The process of iteration 4 results group center as in table 8.

Table 8. Group center of iteration 4

Group	GPA	Credits	Final Project
1	2,48	114,20	0,47
2	2,22	50,18	0,32
3	3,10	147,66	0,76

The objective function is 670.77 so that the objective function changes as follows:
Old distance = 671.97
New distance = 670.77
Changes in objective function = | new distance - old distance |
= | 670.77 - 671.97 |
= 1.20

6. Iteration 5
The process of iteration 5 results group center as in table 9.

Table 9. Group center of iteration 5

Group	GPA	Credits	Final Project
1	2,50	112,25	0,47
2	2,16	46,90	0,30
3	3,10	147,66	0,76

The objective function is 670.00 so that the objective function changes as follows:
Old distance = 670.77
New distance = 670.00
Changes in objective function = | new distance - old distance |
= | 670,00 - 670,77 |
= 0.77

7. Iteration 6
The proces of iteration 6 results group center as in table10.

Table 10. Group center of iteration 6

Group	GPA	Credits	Final Project
1	2,43	109,19	0,47
2	2,17	43,00	0,28
3	3,11	147,41	0,76

The objective function is 671,40 so that the objective function changes as follows:
Long distance = 670.00
New distance = 671.40
Changes in objective function = | new distance - long distance |
= | 671,40 - 670.00 |
= 1.40

8. Iteration 7
The process of iteration 7 results group center as in table 11.

Table 11. Group center of iteration 7

Group	GPA	Credits	Final Project
1	2,45	106,00	0,47
2	2,10	38,63	0,25
3	3,10	147,14	0,75

The objective function is 677.60 so that the objective function changes as follows:
Old distance = 671.40
New Distance = 677.60
Changes in objective function = | new distance - old distance |
= | 677,60 - 671,40 |
= 6.21

9. Iteration 8
The procss of iteration 8 results group center as in table 12.

Table 12. Group center of iteration 8

Group	GPA	Credits	Final Project
1	2,42	104,60	0,47
2	2,10	38,63	0,25
3	3,09	146,86	0,75

The objective function is equal to 681,12 so that the change of objective function as follows:
Old distance = 677,60
New distance = 681.12
Changes in objective function = | new distance - old distance |
= | 681,12 - 677,60 |
= 3.51

10. Iteration 9
The process of iteration 9 results group center as in table 13.

Table 13. Group center of iteration 9

Group	GPA	Credits	Final Project
1	2,42	104,60	0,47
2	2,10	38,63	0,25
3	3,09	146,86	0,75

The objective function is equal to 681,12 so that the change of objective function as follows:
Long distance = 681.12
New distance = 681.12
Changes in objective function = | new distance - long distance |
= | 681,12 - 681,12 |
= 0.00

At the 9th iteration, the value of the object function change of 0.00 is below the threshold value of 0.1 or there is no group movement that the iteration is stopped.

3. RESULT AND DISCUSSION

K-Means algorithm used in a set of Students data were divided into 3 groups through several stages ranging from initialization to iteration 9 which resulted in group movements by finding the

value of object function changes up to 0.00 that was below the threshold value set ie 0.1 or no group movement so that the iteration was stopped. Group movement can be seen in table 2 and table 4 with the number of different group members due to the transfer of data to other groups. Figure 1 shows the graph of each stage on the change of object function performed.

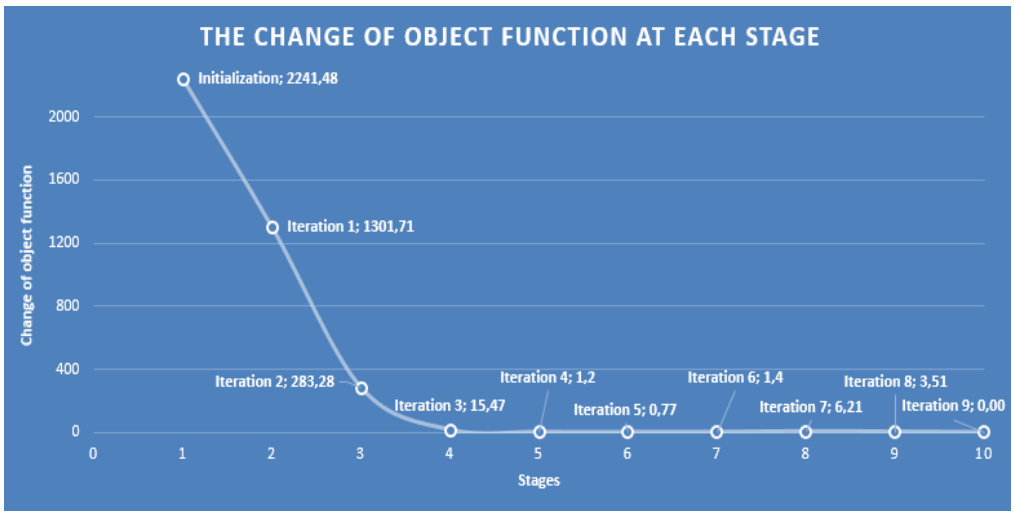

Figure 1. The Change of Object Function at 10 stages

Figure 1 shows a graph of the value of the object function change at the 10 stages performed starting from initialization to the 9th iteration, the graph indicates an increase in value at the initialization stage. To clarify the image from the stages in the 3rd to 9th iteration can be seen in figure 2.

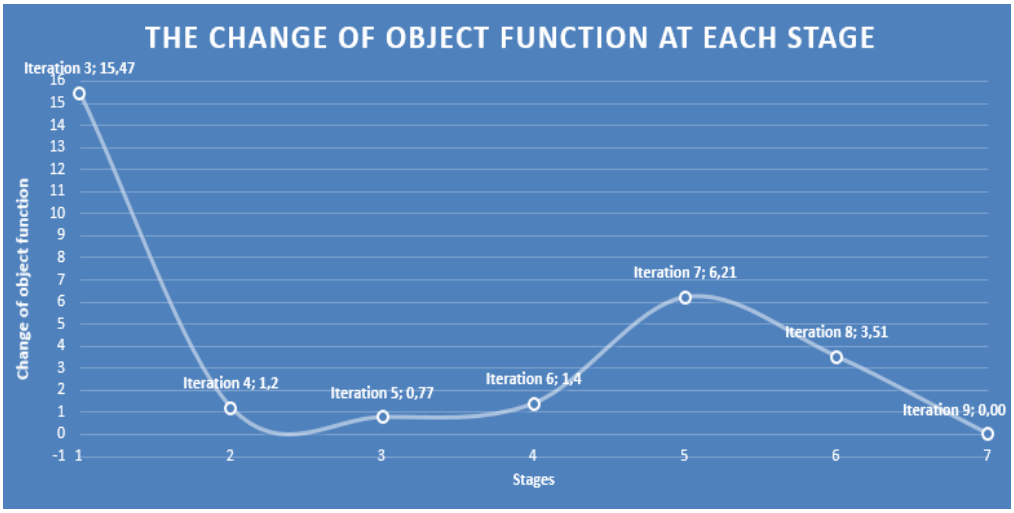

Figure 2. The Change of Object function at last 7 stages

Figure 2 shows the graphs up and down from the change in values occurring in the last 7 stages ie from iteration 3 to iteration 9, the iteration continues to do until it can produce a value below the threshold value as a sign that the iteration can be stopped.

4. CONCLUSION

The conclusion that can be drawn from the use of K-Means algorithm to the set of student data in classifying the academic status of students as follows :

1. K-Means algorithm can be used to classify students' academic status based on GPA, number of credits taken, and final project referenc by applying stages and provisions on K-Means algorithm ie using iteration.
2. Iteration is continuously done as long as the value of the object function is still above the given threshold value of 0.1 or the group movement still occurs. At the 9th iteration, it can be terminated because the it is below the threshold value of 0.00 and no more group movement.
3. The development on the same data can be conducted by using other clustering methods such as hierarchical clustering or Self-Organizing Map so the difference can be reognize and it can be used as other research development.

5. SUGGESTION

Further research may use the same data yet with different amount of data and the number of groupings so it can be known the results of the grouping both in terms of the value of the object function changes and the number of iterations performed.

REFERENCES

[1]Oyelade, O.J, Oladipupo, O.O., dan Obagbuwa, I.C., 2010, Application of k-Means Clustering algorithm for prediction of Students' Academic Performance, *International Journal of Computer Science and Information Security (IJCSIS)*, no 1, vol 7, hal 292-295.

[2] Rosmiati dan Lili, R., 2016, Aplikasi Berbasis Fuzzy C-Means Dalam Penentuan Predikat Kelulusan Mahasiswa, *Jurnal Ilmiah Ilmu Komputer*, no 2, vol 2, hal 1-9.

[3] yulius, P. dan Sitti, N. A., 2017, Pengelompokan Tingkat Kelulusan Mahasiswa Menggunakan Algoritma K-Means, *Seminar Nasional APTIKOM (SEMNASTIKOM)*, Jayapura, November 3.

[4] Harwati, Ardita, P.A., dan Febriana, A.W., 2015, Mapping Student's Performance Based on Data Mining Approach (A Case Study), *Agriculture and Agricultural Science Procedia 3*, hal 173 – 177.

[5] Lili, R., dan Sam'ani, 2016, Pemodelan K-Means Pada Penentuan Predikat Kelulusan Mahasiswa STMIK Palangka Raya, *Jurnal Saintekom*, no 1, vol 6, hal 1-15.

[6] The Regulation of the Minister of Research, Technology and Higher Education of the Republic of Indonesia no.44 year 2015 on National Standards of Higher Education, 2015, Jakarta

[7] Eko, P., 2012, *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*, Andi, Yogyakarta.