Advances in Engineering Research, volume 196
International Joint Conference on Science and Engineering (IJCSE 2020)

Application of K-Means Algorithm for Clustering Student's Computer Programming Performance in Automatic Programming Assessment Tool

Anita Qoiriah[1*], Rina Harimurti[1], Andi Iwan Nurhidayat[1], Asmunin[1]

[1]*Jurusan Teknik Informatika Fakultas Teknik,Universitas Negeri Surabaya, Surabaya, Indonesia*
**Email: anitaqoiriah@unesa.ac.id*

ABSTRACT
Programming is a course that is considered quite difficult for most students. Students are required to have abilities in all processes. Computer programming skills require a lot of practice through lab work assignments. Managing and assessing results of a student's lab work assignment is complex and quite time-consuming task. The availability of automatic programming assessment tools to receive the results of lab work assignments and automatically correct and assess can ease the task of a lecturer. Grouping students according to their level of performance, makes it easy for lecturers to monitor student performance levels and can provide learning according to students' abilities. Grouping was done using K-Means clustering method. Data was score obtained from the lab work assignment of the Automatic Programming Assessment Tool. From the results of clustering, there were 3 groups of students based on their abilities, namely 16 people in the medium ability group, 11 people were students with high ability and 14 people were students whose programming abilities were still lacking.

Keywords: *programming assessment tool, programming, clustering, K-Means*

1. INTRODUCTION

Programming is a fundamental science that students taking degree in computer sciences must study. A significant skill must be possessed by student of the IT department to evaluate and to execute programming languages.

Programming is a course that most students considered rather hard. Programming is an aptitude to level learners before they reach greater concentrations to comprehend the basic abilities. For instance, students must gradually learn syntax and then sémantics structure.

On the other hand, Programming requires some skills. The programming method begins by converting the issue into an algorithm and translating the algorithm into a program code. The hardest part is the application of the algorithm specifications. The right algorithm will produce a program that is as expected. Consequently, students are required to have the ability in all processes; analyze problems, design algorithms, translate algorithms into program code and write program code with the correct syntax. [7].

Computer programming skills require a lot of practice through lab work assignments. Managing and assessing results of student's lab work assignments is very complex and time-consuming. The availability of automatic programming assessment tools to receive grade of lab work assignments, correcting and giving an assessment automatically can ease the task of a lecturer to evaluate student's lab work assignments. The integration of such tools into the Learning Management System is an important feature to improve its performance.

Performance evaluation is one of the basics to monitor the progress of student performance. Grouping students according to their level of performance makes it easy for lecturers to monitor student performance levels and can provide learning according to the abilities of students in these groups. Each group can have different treatment according to what is needed. Grouping can be done by clustering based on the similarity of the performance of each student.

With this grouping, the assignment of practical work and monitoring student performance can be more targeted. In the traditional grouping, students are grouped based on their average scores. In this way, it is difficult to get a comprehensive view of the state of student performance. [6].

The grouping was performed using the K-Means clustering technique through data mining analysis. In this procedure the information was divided into multiple clusters on the basis that the data were similar, so that data with same features were divided into single cluster, while distinct features grouped into another cluster with same features.[5]

This research discussed how the results of the assessment of the automatic programming assessment tool used to classify students based on their performance using the K-Means method.

Automatic Programming Assessment Tools is a tool used to examine programming assignments made by a student. This tool must be able to display lab work assignment, live coding, upload program code, evaluate the program code that has been uploaded, compile, display the compilation results, and do the grading. Besides, this tool can also display scores and rankings from all lab work assignment participants who have uploaded the program code.

In Rina's research, the assessment features developed in the Automatic Programming Assessment Tools referred to the weight of the questions and the correct number of results matching the answer key. The weight of the problem was directly proportional to the level of difficulty and complexity. The more the number of answer keys in the form of input and output, the smaller the percentage of values. [1]

M.Novák's research proposed the integration of plagiarism checking tools namely jPlag and PMD which were used to read source code into a coherent system connected with LMS (Moodle) and IDE. The goal was that sending tasks and checking could be done easily in one system. The system was used in programming courses, object-oriented programming and Java technology [2]. Plagiarism checking was very important, especially in programming tasks.

From the values generated by the automated programming assessment tools, students could be grouped according to their performance. K-means clustering was widely used in educational research. Fan in his research used the K-means algorithm to analyze student data by making improvements to K-means to get better student grouping results. In the proposed algorithm, a method based on network density was used to remove the first outlier. Second, a new method was used to produce the initial cluster center to replace random determinations [9].

Aldi's research used the K-mean method to cluster lecturers' performance whether it was good or not good. The basis used in the assessment used a tri-rhythm basis for the tertiary institution. In assessing performance, data was needed related to student satisfaction with the lecturer. The data used in this study were student satisfaction data. In collecting data using a questionnaire from the Quality Assurance Agency. Variable used were (1) lecturer reliability; (2) responsiveness; (3) guarantees; and (4) empathy. The results of the study were used to improve the performance of lecturers in teaching to improve student satisfaction index [4]

2. METHOD

This chapter consisted of three parts. Section 1 described the data used in this study, section 2 on general methodologies includes steps and section 3 on the clustering methods used in this study.

2.1 Dataset Details

The data in this study were collected from 41 first semester students of the Department of Informatics. Students were participants in the Basic Programming course. The data collected was the result of an Automatic Programming Assessment Tools assessment from a basic programming lab work that uses C ++ Language. The data was collected from a score of 10 lab work assignments that must be completed by each student.

Figure 1 is an automatic programming tool and Figure 2 is an example of a lab work assignment made by a lecturer.

Figure 1. Automatic Programming Assessment Tools

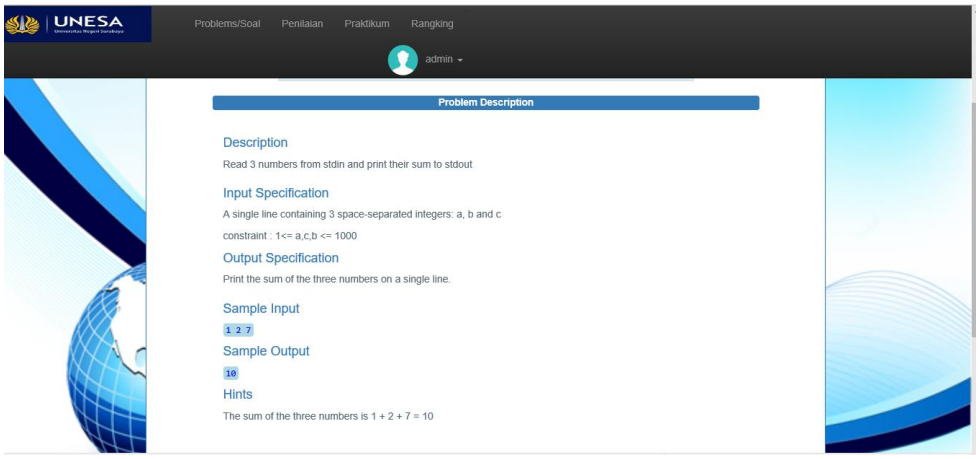

Figure 2. The form of lab work assignment

2.2 K-Means Clustering

K-Means aimed to minimize the Sum of Squared Error (SSE) between data objects with many k-centroids. K centroids in this study amounted to 3, they were a group with low, medium and high programming abilities. The steps for the K-Means algorithm [8]:

1. From the data set to be clustered many k objects are chosen randomly as the initial centroid, cj.
2. Every object that is not centroid is inserted into the nearest cluster based on the Euclidean distance measure: || xi (j) - cj ||2 .
3. Each centroid is updated based on the average of the objects in each cluster
4. Iterate for the second and third steps until all centroids are converging and stable, where all the centroids produced in the current iteration are the same as all the centroids produced previously.

Several distance measurements were used as a measure of similarity of data, one of which is the Euclid distance. The Sum of Squared Error (SSE) as in Equation 1:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad \text{.....(1)}$$

Where *J* is the objective function, *k* is the number of clusters that is 3, *n* is the number of data objects that is 41, *xi* is the value of the data object to *i*, *cj* is the centroid cluster *j*, $\| x_i^{(j)} - c_j \|^2$ is a distance function.

K-Means Algorithm

K-MEANS $(\mathbf{D}, k, \epsilon)$:
$t = 0$
Randomly initialize k centroids: $\mu_1^t, \mu_2^t, \ldots, \mu_k^t \in \mathbb{R}^d$
repeat
  $t \leftarrow t + 1$
  $C_j \leftarrow \emptyset$ for all $j = 1, \cdots, k$
  // Cluster Assignment Step
  foreach $\mathbf{x}_j \in \mathbf{D}$ **do**
    $j^* \leftarrow \arg\min_j \left\{ \left\| \mathbf{x}_j - \mu_i^t \right\|^2 \right\}$ // Assign $\mathbf{x}_j$ to closest centroid
    $C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$
  // Centroid Update Step
  foreach $i = 1$ *to* k **do**
    $\mu_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$
until $\sum_{i=1}^{k} \left\| \mu_i^t - \mu_i^{t-1} \right\|^2 \leq \epsilon$

Figure 3. K-Means Clustering Algorithm

3. RESULT AND DISCUSSION

3.1 Input Data

The test data used was the value of 10 practical tasks generated by the Automatic programming assessment tool. The practicum participants were 41 people. So that the total data was 410 values. Furthermore, the data would be processed into 3 clusters. The cluster described groups of students according to their performance, namely groups that had less, medium and high abilities. Input data was described in graphical form as shown in Figure 4. The x-axis was student and the y-axis showed the grade of lab work assignments. Each lab work assignment was depicted in a different color.

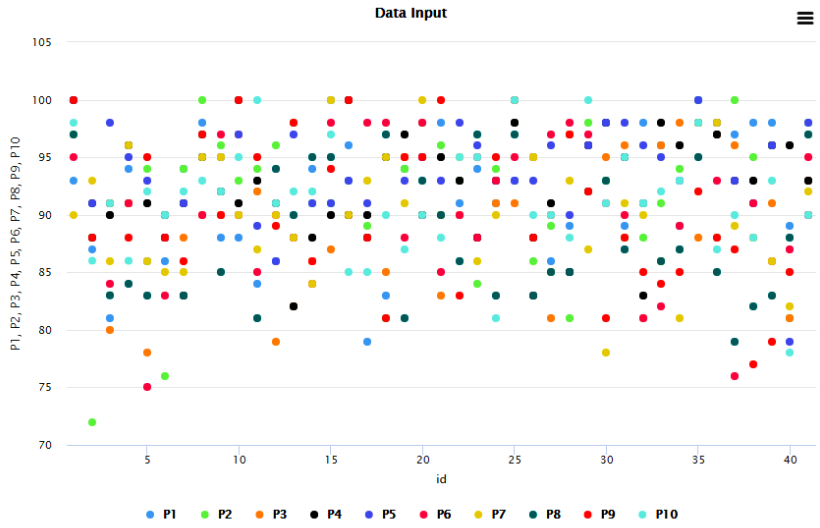

Figure 4. Graphics data input

3.2 Centroid Data

In this study, the data was grouped into 3 groups with low, medium and high programming abilities. In applying the K-means algorithm, these 3 groups were the value of the number of clusters to be created. As an initial centroid, 10 practicum grades were taken from 3 students randomly. Furthermore, from data not centroid, we looked for the minimum distance from the initial centroid. Suppose that an initial centroid is randomly obtained as follows:

Cluster	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
1	87	72	88	91	91	88	93	88	88	86
2	88	98	82	82	97	88	88	90	98	92
3	89	88	96	95	98	90	91	87	88	95

The next process was to update the centroid value based on the average value of the object minimum values. For the first data object, the distance to the initial centroid will be searched, so that it will be obtained as follows:

No	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
1	93	97	100	100	100	95	90	97	100	98

By using the Euclidean distance formula, we obtained the distance to cluster 1 was 37.34, the distance to cluster 2 was 28.72, the distance to cluster 3 was 20.52. The minimum value obtained was the distance to cluster 3, which was 20.52. So the value in cluster 3 was updated to be the midpoint between the centroid in cluster 3 and the data from the object. The centroid value would change to be as follows.

Cluster	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
1	87	72	88	91	91	88	93	88	88	86
2	88	98	82	82	97	88	88	90	98	92
3	91	93	98	97.5	99	93	91	92	94	97

The same process was performed for all subsequent data objects. And the process was repeated until a stable centroid was produced, i.e. the centroid value now and then remains the same. A stable midpoint or centroid value was generated as can be seen as in Table 1.

Table 1. Centroid Data

Attribute	cluster_0	cluster_1	cluster_2
P1	93.375	95.182	87.571
P2	92.938	96.545	86.500
P3	93.062	91.364	84.571
P4	94.312	94.182	89.214
P5	94.312	95.545	91.929
P6	88	94.636	88.500
P7	89.438	93.818	88.286
P8	84.250	95.727	89.571
P9	86.812	96.364	89.714
P10	90.750	94	89.357

3.3 Clustering

By using this centroid, the data would be grouped into 3 clusters. Clustering was determined based on the minimum distance between objects and centroids. The process was repeated by determining the new centroid based on clustering results. The loop was stopped when the same clustering was obtained as before. Clustering results obtained as shown in Figure 5, where cluster 0

contained 16 items or 39% students with moderate ability, cluster 1 as many as 11 items or 27% students with high ability and cluster 2 as many as 14 people or 34% students with lacking programming ability. By looking at the percentage of each group, it appeared that the amount in each group was almost balanced. The results of grouping data with the K-Means algorithm can be described as shown in Figure 6.

Cluster Model

```
Cluster 0: 16 items
Cluster 1: 11 items
Cluster 2: 14 items
Total number of items: 41
```

Figure 5. K-Means Clustering Result

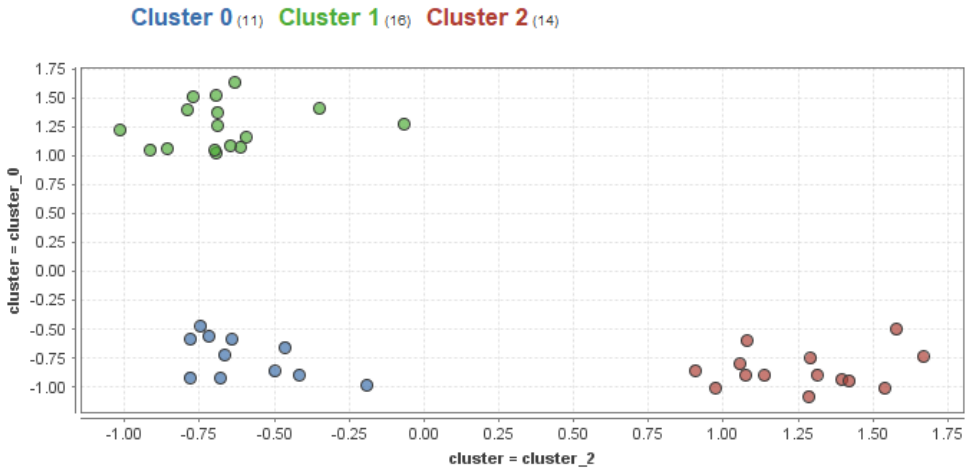

Figure 6. K-Means Scatter Plot

4. CONCLUSION

The K-Means clustering method could be used to group students in programming courses based on their ability. Data was the grade of lab work assignment from the Automatic Programming Assessment Tools. The results of clustering could classify students into 3 clusters; 16 people were in moderate ability groups, 11 students with high abilities and 14 students with programming abilities still lacking.

REFERENCES

[1] H. Rina, Q. Anita, N. Andi Iwan, Asmunin, "Pengembangan Fitur Penilaian dan Perangkingan pada Automatic Programming Assessment Tools", JIEET: Volume 02 Nomor 02, 2018

[2] M. Novák, M. Biňas "Automated Testing of Case Studies in Programming Courses" ICETA 2011 • 9th IEEE International Conference on Emerging eLearning Technologies and Applications Slovakia, October 27-28, 2011

[3] Mohammed J. Zaki, Wagner Meira, Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, May 2014.

[4] N. Aldi, M. Much Aziz, K Miranita "Application of K-Means Algorithm for Clustering Lecturer Based On Assessment of Student Satisfaction Index" Techno.COM, Vol. 16, No. 1: pp 17-24, February 2017

[5] Ong Johan Oscar, "Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University", Jurnal Ilmiah Teknik Industri, Vol. 12, No. 1, Juni 2013

[6] Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance" (IJCSIS) International Journal of Computer Science and Information Security Vol. 7, No. 1, 2010,

[7] R. Masura, S. Shahrina, L. Rodziah, and F. Noor, "Major problems in basic programming that influence student performance" Procedia-Social and Behavioral Sciences 59, pp 287 – 296, Elsevier Ltd., 2012.

[8] Suyanto, Machine Learning Tingkat Dasar Dan Lanjut, Informatika Bandung, Nopember 2018.

[9] Zhongxiang Fan, Yan Sun and Hong Luo "Clustering of College Students Based on Improved K-means Algorithm" Journal of Computers Vol. 28, No. 6, , pp. 195-203, 2017.