IMPLEMENTATION OF K-MEANS CLUSTERING METHOD FOR ELECTRONIC LEARNING MODEL

Herlina Latipa Sari¹, Dewi Suranti², Leni Natalia Zulita³

¹, ², ³ Teknik Infromatika Fakultas Ilmu Komputer Universitas Dehasen Bengkulu Jalan Meranti Raya No.32 Sawah Lebar Kota Bengkulu

Email:¹herlinalatipasari@gmail.com, ²Mrs.dewisuranti@gmail.com, ³Natalia.af25@gmail.com

Abstract - Teaching and Learning process at SMK Negeri 2 Bengkulu Tengah has applied e-learning system for teachers and students. The e-learning was based on the classification of normative, productive, and adaptive subjects. SMK Negeri 2 Bengkulu Tengah consisted of 394 students and 60 teachers with 16 subjects. The record of e-learning database was used in this research to observe students’ activity pattern in attending class. KMeans algorithm in this research was used to classify students’ learning activities using e-learning, so that it was obtained cluster of students’ activity and improvement of student’s ability. Implementation of K-Means Clustering method for electronic learning model at SMK Negeri 2 Bengkulu Tengah was conducted by observing 10 students’ activities, namely participation of students in the classroom, submit assignment, view assignment, add discussion, view discussion, add comment, download course materials, view article, view test, and submit test. In the e-learning model, the testing was conducted toward 10 students that yielded 2 clusters of membership data (C1 and C2). Cluster 1: with membership percentage of 70% and it consisted of 6 members, namely 1112438 Anggi Julian, 1112439 Anis Maulita, 1112441 Ardi Febriansyah, 1112452 Berlian Sinurat, 1112460 Dewi Anugrah Anwar and 1112467 Eka Tri Oktavia Sari. Cluster 2: with membership percentage of 30% and it consisted of 4 members, namely 1112463 Dosita Afriyani, 1112471 Erda Novita, 1112474 Eskardi and 1112477 Fachrur Rozi.

Keywords: Clustering, Learning Model, K-Means Algorithm

1. Introduction
National education goal is to educate nations. It can’t be separated from the role of teacher. In order to improve the quality of education, the development of curriculum, learning innovation, and education facilities contentment are needed. To improve students’ learning achievement, teacher should make innovative learning that support the students learn optimally, either in self-learning and in class learning. In an attempt to improve the quality of education, teachers are required to make learning more effective, innovative and fun [2]. In learning process, the role of teachers is as motivator and facilitator. In consequence of learning paradigm change from instructional based learning into constructional based learning, the teachers should capable to design learning that activate students. To make more effective, innovative, and fun, the teachers can utilize all school resources, either human resources, facilities and infrastructures or other resources [9]. The survey result of this research showed that Sekolah Menengah Kejuruan (SMK) Negeri 2 Bengkulu Tengah has had computer network and already utilize Information and communication Technology (ICT) by utilize-learning in implementing learning into teaching and learning process. It was a step forward that capable to improve students’ learning motivation and also to meet the shortage of meeting in classroom due to
the vocational curriculum that require field study, so that the students can conduct self learning without depend on teacher and books [8]. E-learning is learning media technology that utilize either electronic media or software [9]. E-learning can be developed by using Learning Management System (LMS) called Moodle. Some facilities provided by Moodle includereading module, assignment module, chat module, forum module, option module, and quize module[2]. Once the pattern obtained, analyzing each pattern of subject taken by students was performed, so that it can describe the subject taken by students based on students’ participation interest in following the lessons. Fuzzy K-Means algorithm in this research was used to cluster teaching and learning activities between teacher and students so that it resulted informations about activities group attended by students in e-learning. Beside that, k-means algorithm is also versatile which means that it easy to modify steps in the algorithm, such as in initialization of function to calculate distance and also criterion on stopping iteration [7].

2. Literature Review

Clustering is about grouping of records, observation and forming a class of objects that have similarities. Cluster is a set of records which have similarities to records in another cluster. Clustering is different from the classification that the absence of the target variable in clustering[7]. Clustering plays an important role in data mining applications, such as the exploration of scientific data, accessing information and text mining, spatial database applications, and web analysis[4]. Clustering has two methods, namely hierarchical clustering method and non-hierarchical clustering method. Hierarchical method is used when desired number of groups is not yet known, while the non-hierarchical method is used when the number of groups desired predetermined. Fuzzy K-Means algorithm belongs to the group of hierarchical clustering methods [3]. Fuzzy K-means is a popular cluster algorithm which is used in the process of clustering data sets because of its simplicity[6]. However, this algorithm can also be used for the establishment of a database cluster that its attributes derived from different types, by altering these attributes into the index of similarity or dissimilarity [10]. In the K-Means algorithm cluster number K are already predetermined, object with particular attribute are grouped within cluster K. Each cluster contains cluster center (centroid) point and members of a cluster is selected in accordance with distance from cluster center point [1]. K-Means method belongs to partitioning clustering algorithm [11]. Steps for kmeans clustering are as follow.

a. Decide the number of cluster K
b. Initialization of the cluster center (centroid) can be conducted by using various ways. However, the most frequently used is by using random way. Clusters centers are assigned by random numbers.
c. Allocate all data/objects to the closest cluster. Determination of closeness of two objects are determined based on the distance of two objects. To calculate the distance of all data to each centroid point, Euclidean Distance theory is used that is formulated as follow.

\[ D_{ij} = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 + \ldots + (X_{in} - X_{jn})^2} \]

Where: \( D_{ij} \) = distance of ith data to cluster center j \( X_{ki} \) = the ith on the kth data atribut \( X_{kj} \) = the jth center point on the kth data atribut
d. Recalculate centroid with current cluster membership. Centroid is an average (mean) of all data/objects within particular cluster. If desired, the median of this cluster can also be used. Thus the mean is not the only one parameter that can be used.
e. Reassign each object by using new cluster center, if the cluster doesn’t change, then clustering process finished otherwise repeat step 3 until there is no change for each cluster.

3. Research Methodology

The methodology used in this research illustrates the steps taken to perform the analysis of the implementation of K-Means Clustering Algorithm for electronic based learning in the form of E-learning application at SMK Negeri 2 Bengkulu Tengah. Figure 2 shows research framework started by defining the scope of the problems to the testing of K-Means Clustering algorithm. Once direct observation done, literature study was conducted by comparing some previous research. Partitioning data aims to classify data that has same characteristics into one same cluster and data that has different characteristics into one other group [1]. After defining the scope of the problem, then analyzing the problem and determine the expected goals was performed followed by studying the literature about K-Means Clustering Algorithm and the results of calculations and grouping so that it yielded grouping of student activity in learning activities in the E-Learning application.

The data in this research were drawn from the database that is contained in the elearning application in the form of teacher login data and student login data that contains student activity in participating lessons given in the teacher [12].

There are three group of subject in the elearning, namely normative, adaptive, and productive. Activities type that exist in elearning are shown in Table 1.

| No | Students Activities E-Learning Module |
|----|----------------------------------------|
| 1  | Students participation in classroom    |
| 2  | Submit Assignment                      |
| 3  | View Assignment                        |
| 4  | Add Discussion                         |
| 5  | View Discussion                        |
| 6  | Add Comment                            |
| 7  | Download Course Material                |
| 8  | View Article                           |
| 9  | View Text                              |
| 10 | Attempt Test                           |

| No | Students Activities E-Learning Module |
|----|----------------------------------------|
| 1  | course                                 |
| 2  | Assignment                             |
| 3  | Assignment                             |
| 4  | Forum                                  |
| 5  | Forum                                  |
| 6  | Forum                                  |
| 7  | Resource                               |
| 8  | Blog                                   |
| 9  | Quiz                                   |
| 10 | Quiz                                   |

4. Discussion

K-Means algorithm is used in clustering dataset due to simplicity of its algorithm to process big dataset, in addition the K-means algorithm is a partitioning clustering that separates data into different groups. Through partitioning iteratively it can minimize the average distance of each data into clusters [12]. The implementation of K-Means Clustering algorithm in this research was conducted by simulating students activities with consist of 396 students. In this study, 10 students were taken as samples by observing the activities of 10 students in applying E-Learning with 10 activities that students activities with NIS 1112438 named Anggi Julian are 60 times entering classroom, 18 times submitting assignments, 20 times viewing assignments, never adding discussion, never viewing discussion, never adding comments, 25 times downloading course materials, never viewing articles, 25 times viewing test, and 25 submitting tests.

From the total of students activities data shown in Table 2, the steps in implementing K-Means algorithm are as follow.
a. Number of desired clusters are 2 clusters (C1 and C2)
b. Determiniation of Centroid (cluster center point) In this research, second data is used to determine centroid for cluster 1 and fifth data is used for determination centroid for cluster 2. These two data are as folow:
Cluster 1 : 79, 11, 19, 0, 0, 29, 0, 35, 30
Cluster 2 : 45, 0, 21, 0, 0, 8, 26, 1, 26, 22
c. Calculate the distance for every data to its centroid

Iteration 1: Calculation of distance for 1st data:
$$D_1 = \sqrt{(79-60)^2 + (11-18)^2 + (19-20)^2 + (0-0)^2 + (0-0)^2 + (29-30)^2 + (0-0)^2} = 21.84$$
$$D_2 = \sqrt{(45-60)^2 + (0-18)^2 + (21-20)^2 + (8-25)^2 + (26-30)^2 + (12-25)^2} = 19.95$$

d. The same calculation is then conducted toward 2nd to 10th data, once calculating the distance of centroid for all data, the distance matrix is then obtained as follow. D1:

| Cluster | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|---------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| C1      | 21.84 | 0   | 16.46 | 41.06 | 0   | 33.60 | 21.93 | 32.97 | 49.44 | 61.36 |
| C2      | 19.93 | 37.19 | 32.47 | 24.10 | 40.78 | 12.47 | 19.39 | 13.19 | 30.66 | 40.71 |

Every column in the matrix represents the distance value of data to centroid. First row in the matrix represents distance value of data to the first centroid (C1) : 79, 11, 19, 0, 0, 29, 0, 35, 30 and the second row in the matrix represents distance value of data to second cluster (C2) : 45, 0, 21, 0, 0, 8, 26, 1, 26, 22. In the cluster 1 (C1), the distance for the first activity (participation of students in classroom) to the centroid point is 21.84 and in the cluster 2 (C2) the distance value for the first activity (participation of students in classroom) to the centroid point is 19.95. In the cluster 1 (C1), the distance value for the second activity (submitting task) to the centroid point is 0, and in the second cluster (C2) the distance value for the second activity (submitting task) to the centroid point is 37.19.
e. Data Grouping The calculated distances are then compared to select the closest distance between data and centroid, this distance show that the data are in a cluster with closest centroid. Here will be shown matrix data of cluster, value 1 means that the data are in the cluster dan value 0 means that the data are not in the cluster. G1 =

| Cluster | Activity | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---------|----------|---|---|---|---|---|---|---|---|---|----|
| C1      | 0        | 1 | 1 | 0 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| C2      | 1        | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1  |

In the cluster 1 the value of activity 1 (A1) is 21.84, in the cluster 2 is 19.95, it means that C1A1 < C2A1 so that the smaller value close to centroid is assigned value 1 whereareas C1A1 is assigned
value 0. The same process was also conducted to C1A2 and C2A2 to C1A10 and C2A10. In the data grouping in G1 for the number of cluster 1 (C1) contains 3 members (2,3,5) and the number of cluster 2 (C2) contains 7 members (1,4,6,7,8,9,10).

g. Determination New Centroid Once the members of each cluster obtained, then a new centroid are calculated based on data from each cluster member in accordance with the formula of the center of the cluster members. So that we get the following calculation: Cluster 1 (C1) obtained 3 members with calculation of new centroid is as follow.

g. Data Grouping As in the first iteration, the calculated distances are then compared to select the closest distance between data and centroid, this distance show that the data are in a cluster with closest centroid. Here will be shown matrix data of cluster, value 1 means that the data are in the cluster and value 0 means that the data are not in the cluster. G2 = In the cluster 1 the value of activity 1 (A1) is 13.35, in the cluster 2 is 18.84, it means that C1A1 < C2A1 so that the smaller value close to centroid is assigned value 1 whereas C1A1 is assigned value 0. The same process was also conducted to C1A2 and C2A2 to C1A10 and C2A10. In the data grouping in G1 for the number of cluster 1 (C1) contains 5 members (1,2,3,6,7) and the number of cluster 2 (C2) contains 5 members (4,5,8,9,10). Based on the calculation in the first and second iteration with G1 and G2 grouping, the equal number of cluster 1 (C1) and cluster 2 (C2) has not been reached so that it needs to repeat the calculation in determining new centroid. In this research, the stabil clustering result and convergent was reached at iteration 5 with the matrix of students activities G5 as follow.

h. Simulation Result of Clustering Analysis Based on the calculation K-Means clustering algorithm using simulation can be interpreted as follow. Cluster 1: with membership percentage of 70% and it consisted of 6 members Cluster 2: with membership percentage of 30% and it consisted of 4 members h. Implementation Process of K-Means Clustering Algorithm using RapidMiner In implementing K-Means algorithm using RapidMiner the first step is by importing excel data “Aktifitas Siswa”

| Activity                      | Min | Max  | Average | Deviation |
|-------------------------------|-----|------|---------|-----------|
| Entering Class                | 48  | 79   | 53.400  | 14.238    |
| Submit Assignment             | 6   | 18   | 11.300  | 3.917     |
| Adding Discussion             | 19  | 58   | 27.600  | 11.316    |
| Viewing Discussion            | 0   | 0    | 0       | 0         |
| Adding Comment                | 0   | 8    | 0.900   | 2.514     |
| Downloading Course Material   | 7   | 40   | 24.600  | 9.228     |
| Viewing Article               | 0   | 1    | 0.400   | 0.516     |
| Viewing Test                  | 19  | 42   | 30.200  | 7.524     |
| Attempting Test               | 15  | 40   | 25.700  | 7.072     |

After running K-Means Clustering, it will appear the graphic of data clustering using RapidMiner as shown in Figure 7.
5. Conclusion

K-Means algorithm in this research was used to classify students’ learning activities using e-learning, so that it was obtained cluster of students’ activity and improvement of student’s ability. The testing was conducted toward 10 students that yielded 2 clusters of membership data (C1 and C2). Cluster 1: with membership percentage of 70% and it consisted of 6 members, namely 1112438 Anggi Julian, 1112439 Anis Maulita, 1112441 Ardi Febriansyah, 1112452 Berlian Sinurat, 1112460 Dewi Anugrah Anwar and 1112467 Eka Tri Oktavia Sari. Cluster 2: with membership percentage of 30% and it consisted of 4 members, namely 1112463 Dosita Afriyani, 1112471 Erda Novita, 1112474 Eskardi and 1112477 Fachrur Rozi.

REFERENCES

[1] K J. Kim and H Alin“A Recommended System Using GA K-Means Clustering in an Online Shopping Market “, Expert System with Application Vol. 34, PP.1200-1290, 2008
[2] Miyazaki Y & Kurasige K, “Use of Reward Independent Knowledge on Reinforcement Learning for Dynamic Environment”, ICACSIS 2010 Faculty of Computer Science Universitas Indonesia.
[3] Handoyo R, Rumani R, Nasution SM,” Perbandingan Metode Clustering menggunakan Metode Single Linkage dan K-Means pada pengelompokan dokumen” JSM STMIK Mikroskil Vol 15, No.2 Oktober 2014 ISSN 1412-0100
[4] J. Santoso, G.O, Van Albada, B. Nazief, P.M.A. Sloot, “Hierarchical Job Schedulling for Cluster of Workstation”, (Published Conference Proceeding Style in Proc the 6th ASCI Delft Netherland, June 2010.
[5] Celebi, M.E, Kingraul, HA & Vela P.A, “Comparative Study of Efficient Intralization Methods for the K-Means Clustering Algorithm”, Expert System with Application, 2013.
[6] Grossman, David A dan Ophir Frieder, “Information Retrieval Algorithm and Heuristics”, Second Edition Springer, The Netherland, 2004.
[7] Suhendi, “Analisa Penggunaan E-learning untuk meningkatkan kemudahan Mahasiswa dalam pembelajaran”, Seminar Nasional Teknologi Informasi dan Multimedia 2015. STMIK Amikom Yogyakarta, ISSN: 2302-3805, 2015
[8] Surjono HD, “Membangun Course E-Learning Berbasis Moodle” I, UNY press Yogyakarta, 2011
[9] Rokach L, “A Survey of Clustering Algorithm”, in O Malman &L. Rokach (eds). Data Mining and Knowledge Discovery Handbook (2nd ed) springer science + Business Media LLC.
[10] Rivani Admira, “Aplikasi K-Means Cluster untuk Pengelompokkan Provinsi Berdasarkan Produksi Padi, Jagung, Kedelai, dan Kacang Hijau tahun 2009”, Jurnal Mat Stat, Vol.10 No.2 Juli 2010 :122-134. [11] Wijaya AK, “Implementasi Data Mining dengan Algoritma Fuzzy C-Means Studi Kasus Penjualan di UD. Subur Baru”, Fasilkom Udinus, 2014