INITIAL CENTROID OPTIMIZATION IN K-MEANS ALGORITHM FOR EDUCATIONAL DATA MINING

Raga S.H. Istanto, Arinda D. Listikowati, Bekti P. Wibowo, Fitra A. Bachtiar

Brawijaya University, Indonesia, fitra.bachtiar@ub.ac.id

ABSTRACT

Senior High School is a formal educational institution at the level of secondary education that aims to develop students' potential in various disciplines. In the process of development, students are directed to certain interests. In the traditional method, the process of selecting interests is carried out by the teacher which turns out to be difficult, especially in identifying and finding information that is useful to determine student areas of interest. To overcome these problems, the Educational Data Mining (EDM) approach is used. K-Means is one method that can be used in EDM. However, the K-Means method has a disadvantage in choosing initial centroids, because it is slow and less accurate. This study proposed 3 methods for conducting initial centroid selection. Evaluation results obtained in the study using 245 student data showed that the initial centroid using average calculations on data that had been normalized using min-max normalization obtained better results.

Keywords: Initial Centroid, K-Means, Educational Data Mining.

1. INTRODUCTION

In Indonesia, based on government regulation no. 17 of 2010 article 79 (PP, 2010), there is a process for dividing senior high school students into several interests, which are named natural science study programs, social science study programs, and other study programs that are needed by the community. In the traditional method, the process of selecting interests is done by the teacher. The problem with the traditional selection process is that teachers must identify and find useful information on large data manually which is a difficult task to do (Quadril and Kalyankar, 2010). A very promising solution to facilitate the process of selecting interests is to use educational data mining (EDM) (Romero, 2007).

This approach has been widely used to carry out data grouping processes such as the case for selecting student interests. There are many methods in the educational data mining approach to grouping data. K-means is a method of grouping data that is very popular to use because of its simplicity which makes computing faster and uses memory more efficiently (Singh *et al.*, 2011). K-Means is a partitioning clustering method that separates data into different k groups. With the principle of partitioning iteratively, K-Means minimizes the average distance of each data to its cluster. However, the performance of K-means is strongly influenced by the selection of the initial centroid center point (initial centroid) (Mahmud, 2012). K-Means raises the initial center point of the cluster through random selection. If the randomly selected center point of the cluster is approaching the final solution of the cluster center, the k-means algorithm will work faster and uses memory more efficiently without any incorrect clustering results. Conversely, if the initial center point of the cluster is far from the center of the final solution cluster, then it is very likely to cause incorrect clustering results (Cheung, 2003). Poor choice of initial centroid because random selection can cause this method to be slow and inaccurate in grouping data.

Several methods have been introduced to optimize the initial centroid for the K-Means algorithm. Duda and Hart (1973), have discussed a recursive method for initializing the average value obtained from the whole data and randomly generated k times the initial center point. Bradley and Fayyad (1998) propose an algorithm that can optimize the starting point by analyzing data distribution and data density probabilities. Shehroz and Ahmad (2004) introduced a method called the Cluster Center Initialization Algorithm (CCIA) to complete the initialization of the initial center point for K-means. CCIA calculates the average values and standard deviations for all data attributes and then separates the data using a normal curve to a particular partition. CCIA uses K-means and density-based multi-scale data conditions to observe the similarity of data patterns before finding the starting point for K-means.

Based on these explanations, the researcher is interested in proposing the optimization of initial centroid selection by proposing new methods resulting from variations and combinations of methods that already exist. Determination of the initial center point will be done through a series of stages of normalizing data, separating a number of data with normal curves to certain partitions using the vote technique, and initializing the initial center point using the average value of each partition. The results of the proposed optimization will then be evaluated so that it can be shown changes in the level of accuracy, the number of iterations, and data density.

CEUR-WS.org/Vol-2627/short7.pdf

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. PROPOSED METHOD

2.1 Data Pre-processing

The dataset used in this study is a dataset of student grades taken from one school in Indonesia. The raw dataset is 265 data. The dataset has 12 features in the form of 9 subject values ranging from 0 to 100, student id, name and one feature output in the form of classes displayed in IPA (natural science) or IPS (social) class names. Raw data that is separated into several files are put together in one container. The duplicate grades are removed based on the student full name match. Data that has a feature with a value of 0 and/or null is also removed. Finally after pre-processed, the data used in this study were 245 data.

2.2 Proposed Method 1

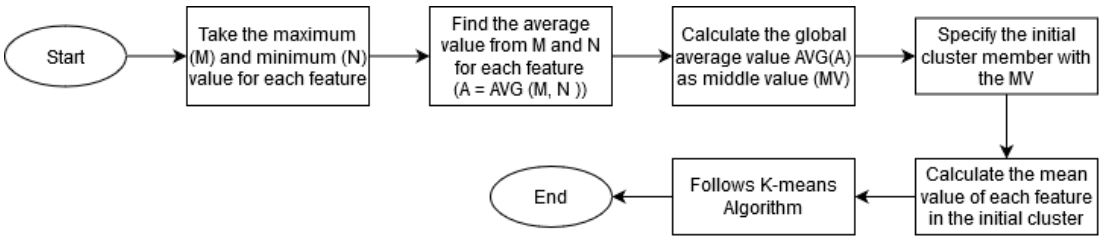

Fig. 1. The framework of proposed method 1

The first proposed optimization for initial centroid selection that can be seen in Figure 1 starts with finding the maximum value called (M) and minimum values (N) for each feature, then find the average value for each feature using M and N values. After that calculate the global average value as the middle value of the dataset to be clustered. The middle value is obtained by finding the average value of the average value of each attribute in the dataset, this value is called the middle value. After the middle value is obtained, then insert each object into the initial cluster using the calculation of the majority of data membership. For example, if an object has data [70,75,71] on all three attributes and the middle value that is calculated is 74 then the value will be [L, H, L] in other words the object will be entered into a cluster member L. After each object is entered into in the initial cluster, the next step is to calculate the average value of each attribute in the cluster to be used as the initial centroid. The next step in clustering follows the K-means algorithm after getting the initial centroid. Assign each data to a specific cluster based on calculation of the distance from each initial centroid where the distance of data from the centroid is minimum using Euclidean distance. Then revive new centroids from each cluster and reassign data based on the distance from the new centroid. Continue to repeat until there are no cluster changes in each data.

2.3 Proposed Method 2

The second proposed optimization of initial centroid selection starts with finding the middle value of the dataset to be clustered. The middle value is obtained by finding the average value of the average value of each attribute in the dataset, this value is called the middle value. After the middle value is obtained, then insert each object into the initial cluster using the calculation of the majority of data membership. For example, the object has data [69, 78, 73] on all three attributes and the middle value that is owned is 74 then the value will be [L, H, L] in other words the object will be entered into the members of cluster L (the process is broadly the same as the first proposed method to get initial group members, see Figure 1 for details). After getting all members in every initial cluster, initial centroids are then randomly selected from each cluster. The next clustering step follows the K-means algorithm after determining the initial centroid.

2.4 Proposed Method 3

The third optimization proposed for initial centroid selection that can be modeled in Figure 2 starts with normalizing the dataset using min-max normalization. After the data is normal, it then determines the initial K-Members randomly. After each object is clustered into a number of initial K, the mean value is calculated for each cluster to be used as the initial centroid. The next clustering step follows the K-Means algorithm until there is no cluster displacement on each object.

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

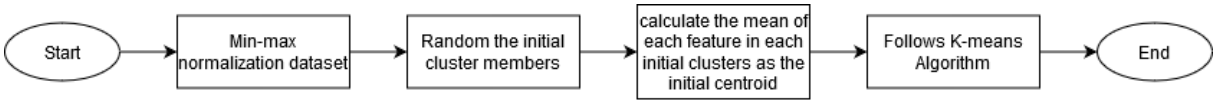

Fig. 2. The framework of proposed method 3

3. ANALYSIS

3.1 Initial Method

At this stage, testing of the K-means method in general is carried out. Testing is done by looking at the value of many iterations, accuracy, and silhouette values. Testing is done by selecting the initial centroid randomly that can be seen in Table 1. Selected the initial centroid by random for the data line [12.28; 120.40; 25.197; 227.87; 5.245]. Tests carried out as many as 5 until the test results can be presented in Table 2.

Table 1. Initial centroid result using k-means.

Experiment	Cluster	AGM	KWN	IND	ING	MTK	SEJ	KES	PJS	PK
1	1	83	82	82	94	79	86	75	80	85
	2	88	77	84	80	78	90	75	82	85
2	1	79	75	80	80	80	89	77	80	80
	2	85	79	82	86	88	86	75	82	75
3	1	86	76	78	80	62	76	75	80	80
	2	78	75	80	80	76	83	75	80	80
4	1	81	80	78	80	75	75	76	82	80
	2	80	75	80	80	75	75	75	80	85
5	1	91	78	78	81	63	85	75	80	80
	2	88	75	76	80	75	75	75	80	80

Table 2. K-means test results.

Sample	Initial Centroid (data-id)	Iteration	Accuracy	Silhouette
1	12;28	7	70.20	-0.283
2	120;40	8	69.80	0.063
3	25;197	18	68.16	0.126
4	227;87	9	68.57	-0.241
5	5;245	17	68.57	0.120
	Average	11.8	69.06	-0.0432

The results of the K-means method show that the number of iterations is at least 7 and the most are 18, the lowest accuracy is 68.16 with the highest accuracy of 70.20, and the lowest silhouette value is -0.283 with the highest silhouette value is 0.126. This can occur due to incorrect selection of initial centroids, or initial centroid much difference from the final centroid. Compared to the first proposed method, regular k-means method provide an average 0.8 more iterations, accuracy is 0.9% better and the silhouette value decreases by 0.169.

3.2 Proposed Method 1

The test scenario performed on the proposed method 1 is to select the initial centroid through the average value in the dataset which has been normalized on the normal curve whose middle limit is the average value of the average

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

minimum and maximum value of each feature or attribute, see Table 3 for details. The middle limit that is obtained is 79.88. Furthermore, each dataset will be categorized into 2 data groups, below the limit and above the limit. After it is divided into 2 groups, the average value of the initial centroid is selected. Initial centroids obtained are [82.39; 75.60; 78.14; 80.49; 73.19; 76.62; 75.25; 79.91; 80.00] for cluster 1, while for cluster 2 [85.76; 78.76; 81.11; 81.04; 78.96; 82.51; 75.43; 80.89; 82.47]. The test results using the proposed initial centroid 1 selection method can be shown in Table 4.

Table 3. Initial centroid result using proposed method 1.

Experiment	Cluster	AGM	KWN	IND	ING	MTK	SEJ	KES	PJS	PK
1	1	82.39	75.60	78.14	80.49	73.19	76.62	75.25	79.91	80.00
	2	85.76	78.76	81.11	81.04	78.96	82.51	75.43	80.89	82.47

Table 4. Proposed method 1 testing result.

Iteration	Accuracy	Silhouette
11	68.16	0.1258

Because it uses the average value to select the initial centroid and the initial group selection uses the vote system, only 1 experiment can be conducted. From the test results, it can be seen that the number of iterations is reduced by 0.8 when compared to the initial centroid random selection method. Accuracy decreased by 0.9%, and silhouette values were better by 0.1690.

3.3 Proposed Method 2

Proposed method 2 randomly selects initial centroids on a normalized dataset on a normal curve whose middle limit is the average value of the average minimum and maximum values of each feature or attribute. The middle limit that is obtained is 79.88. Furthermore, each dataset will be categorized into 2 data groups, below the limit and above the limit. After it is divided into 2, then the initial centroid will be randomly selected by 5 samples that can be shown in Table 5. The results of testing using the proposed method 2 can be presented in Table 6.

Table 5. Initial centroid result using proposed method 2.

Experiment	Cluster	AGM	KWN	IND	ING	MTK	SEJ	KES	PJS	PK
1	1	77	75	81	80	75	75	75	82	90
	2	86	77	81	80	75	93	75	82	80
2	1	84	75	77	81	82	76	75	78	80
	2	90	79	84	80	80	77	76	82	90
3	1	75	75	79	84	75	75	75	80	75
	2	80	75	80	80	79	79	75	82	80
4	1	78	85	79	80	75	80	75	80	75
	2	89	88	82	80	79	90	76	82	85
5	1	79	77	78	81	75	75	78	80	80
	2	88	82	83	84	92	84	75	82	85

The results of testing the initial centroid selection method 2 give the same level of accuracy results with the random selection of initial centroids on the usual K-means method. However, the average iteration 3.8 fewer and silhouette values increased by 0.1516.

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Table 6. Proposed method 2 testing result.

Sample	Initial Centroid (data -id)	Iteration	Accuracy	Silhouette
1	169;162	6	68.57	0.121
2	242;123	7	68.98	0.106
3	37;90	16	68.57	0.121
4	145;244	6	68.98	0.106
5	133;218	5	70.20	0.088
	Average	8	69.06	0.1084

3.4 Proposed Method 3

In the proposed method 3 centroid selection will be done using the average value in the dataset that has been grouped randomly after going through the normalization process. Normalization using min-max normalization. After knowing the normal form in the dataset, then each data is allocated into the initial group randomly, into 2 data groups. From these groups, an average value will be calculated to be used as the initial centroid. Random grouping carried out 5 times, resulting in 5 initial centroid samples that can be seen in table 7 as experimental material. With the result of testing the proposed method 3 can be shown in Table 8.

Table 7. Initial centroid result using proposed method 3.

Experiment	Cluster	AGM	KWN	IND	ING	MTK	SEJ	KES	PJS	PK
1	1	0.40	0.59	0.39	0.18	0.57	0.41	0.04	0.81	0.42
	2	0.38	0.59	0.35	0.18	0.56	0.37	0.04	0.78	0.47
2	1	0.36	0.59	0.37	0.18	0.56	0.40	0.03	0.79	0.45
	2	0.42	0.59	0.37	0.18	0.56	0.39	0.05	0.81	0.45
3	1	0.38	0.60	0.37	0.19	0.57	0.41	0.04	0.82	0.43
	2	0.40	0.57	0.37	0.17	0.56	0.38	0.04	0.78	0.46
4	1	0.40	0.60	0.36	0.16	0.55	0.40	0.04	0.79	0.43
	2	0.37	0.58	0.38	0.19	0.57	0.39	0.03	0.80	0.46
5	1	0.37	0.59	0.39	0.19	0.56	0.42	0.04	0.79	0.45
	2	0.40	0.59	0.35	0.16	0.57	0.37	0.03	0.80	0.44

Table 8. Proposed method 3 testing result.

Sample	Iteration	Accuracy	Silhouette
1	8	73.88	-0.266
2	5	68.16	0.101
3	9	73.88	-0.266
4	12	73.47	0.102
5	6	73.06	-0.259
Average	8	72.49	-0.1176

From the test results when compared with the random selection of initial centroids on the usual K-means method, the proposed method 3 provides an average of 3.8 fewer iterations, accuracy increase 3.43%, and the silhouette value decreases 0.07444. Compared to the first proposed method, proposed method 3 provide an average

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

of 3 fewer iterations, accuracy is 4.33% better and the silhouette value increases by 0.2434. While compared to the second proposed method, the proposed method 3 gives the same iteration average as the proposed method 2, accuracy increase 3.43% and the silhouette value increases by 0.226.

3.5 Experiment Result

From the result of experiments concluded on 245 data using 3 proposed methods for selecting initial centroid with 5 experiments in each proposed method. The result obtained for the least average iteration and the best accuracy is obtained in the third proposed method. However instead, the level of data density in the third proposed method is worse when compared to the 2 other proposed methods. Meanwhile, the best data density is obtained in the first proposed method, but the average iteration and accuracy obtained cannot be as good as in the third proposed method.

4. CONCLUSION

Based on experiments, it is known that from the 3 proposed optimization methods it can be concluded that for the number of iterations and the level of accuracy of the proposed optimization method 3 namely min-max normalization shows the best results, while for the best silhouette value is owned by the first proposed method. In future work, a similar dataset with a larger amount of data can be tested using all proposed methods. Other data sources can also be used to find out other results from all proposed methods.

REFERENCES

Bradley, P.S., and Fayyad, U.M. (1998). Refining initial points for K-means clustering, In: *Proceeding of the 15th International Conference on Machine Learning (ICML'98)*, Shavlik, J. (ed.), 91-99. Morgan Kaufmann: San Francisco.

Cheung, Y.M. (2003). k-Means: A new generalized k-means clustering algorithm. *Pattern Recognition Letters*, 24(15), 2883-2893.

Duda, R.O., and Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. Wiley: New York.

Khan, S.S., and Ahmad, A. (2004). Cluster center initialization algorithm for K-Means clustering. *Pattern Recognition Letters*, 25(11), 1293–1302.

Mahmud, S., Rahman, M., and Akhtar, N. (2012). Improvement of K-means Clustering algorithm with better initial centroids based on weighted average, In: *7th International Conference on Electrical and Computer Engineering*, 647-650. IEEE: New York NY.

PP. (2010). *Peraturan Pemerintah Republik Indonesia nomor 17 tahun 2010 Tentang Pengelolaan dan Penyelenggaraan Pendidikan*. Presiden Republik Indonesia: Indonesia.

Quadril, M.N., and Kalyankar, N.V. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*, 10(2), 2-5.

Romero, V., and Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. *Expert System with Applications*, 33(1), 135-146.

Singh, K., Malik, D., and Sharma, N. (2011). Evolving limitations in K-Means algorithm in data mining and their removal. *International Journal of Computational Engineering & Management* (IJCEM), 12, 105-109.

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).