Chapter 49
Identify Education Quality Based on Islamic Senior High School Data in *Kompetisi Sains Madrasah* Using Fuzzy C-Means Clustering

Dian Candra Rini Novitasari, Abdullah Faqih, Noor Wahyudi, Nurissaidah Ulinnuha, Zakiyatul Ulya, Ali Mustofa, Ahmad Fauzi, Ahmad Hanif Asyhar, Kusaeri, Dwi Rolliawati, and Ahmad Yusuf

Abstract The education system in a country certainly has goals to be achieved. Many attempts were made to achieve the success of an education system that was created. This case causes competition between educational institutions and between students. The results of the competition can also represent schools that need special attention to improving learning systems. The learning system in Indonesia needs evaluation because of educational quality not perfectly. In this research, the fuzzy c-means (FCM) method is used to cluster the quality of existing Kompetisi Sains Madrasah (KSM) participant data to find schools that have good quality education. Next, a cluster evaluation is performed to determine the success of the cluster using silhouettes. Based on 4 trials, that are 3 clusters, 4 clusters, 5 clusters, and 6 clusters, the best results are on 3 clusters with the best silhouette value of 0.6499 with a standard deviation value of 0.2419. In the identification process, the results obtained that SMA has the best education quality in KSM.

49.1 Introduction

Education is one of the conditions for the country's development. The education system in a country has a goal to be achieved [1]. Many attempts were made to achieve the success of the education system. This causes a lot of competition between educational institutions and students [2]. Competition in education usually occurs in a competent competition. Competence is the ability or skill to execute anything without considering competitors, but only knowledge and abilities [2]. Verhoeff,

D. C. R. Novitasari (✉) · N. Wahyudi · N. Ulinnuha · Z. Ulya · A. Mustofa · A. Fauzi ·
A. H. Asyhar · Kusaeri · D. Rolliawati · A. Yusuf
UIN Sunan Ampel Surabaya, Ahmad Yani 117, Surabaya, Indonesia
e-mail: diancrini@uinsby.ac.id

A. Faqih
Kementerian Agama RI, Lapangan Banteng 3-4, Jakarta, Indonesia

Y.-D. Zhang et al. (eds.), *Smart Trends in Computing and Communications: Proceedings of SmartCom 2020*, Smart Innovation, Systems and Technologies 182,
https://doi.org/10.1007/978-981-15-5224-3_49

Lawrence, Fulu, and others suggested that competition in education has excellent benefits because it can increase student motivation and learning, even weaker students can survive by participating in the competition [3].

Indonesia is a developing country that has challenges in competition between nations in the global era to improve the quality and productivity of educated people [4]. One of the efforts made by the Ministry of Religious Affairs is the holding of Kompetisi Sains Madrasah (KSM). KSM is a madrasah science competition that was held in Indonesia. Initially, this competition was intended for students of Madrasah Ibtidaiyah (Islamic Elementary School), Madrasah Tsanawiyah (Islamic Junior High School), and Madrasah Aliyah (Islamic Senior High School). However, since 2016, the KSM has expanded its reach to elementary, junior high school, and senior high school students. KSM competition is held at the district/city provincial and national levels [5].

The competition scores show the results of learning and the ability of students to the learning system that has been applied. Roberta et al., in their research, also stated that the results of a competition held could evaluate the quality of the education system in a country [6]. The results of the competition can also represent schools that need special attention to improving learning systems [7]. The technique to find out schools that need quality improvement can use the clustering method [8]. Clustering is a method for grouping data in multidimensional data based on the similarity of data characteristics [9]. The success rate of the clustering process is at the center of the cluster. The more precise cluster center in representing data differences can provide better clustering results [10].

Elock Emvula Shikalepo has researched clustering on the quality of education in rural Namibian schools. The clustering method used is the Five Cluster Center Principals (CCPs) that have a good result [11]. Lotfi Nadji and Brahim Er-Raha have also studied the clustering method in education. The method used for clustering is the Within-cluster Sum of Square (WSS) method. The study successfully conducted a cluster based on student quality at IBN-ZOHR University. In that research, recommend the fuzzy c-means (FCM) method for clustering to determine the quality of each student. Based on the suggestion of the research, this study will use the FCM method to carry out the clustering process in identifying the quality of education in Indonesia [12].

FCM is a technique for clustering data where at each point that is located in the cluster is determined by the membership value [13–15]. Govindasamy and Velmurugan also researched student performance in a class by comparing several clustering methods such as k-means, k-medoids, and FCM. The results of this research on each method using purity as a cluster evaluation are 0.375, 0.374, and 0.624, and the best results in the research are using the FCM method [16]. Results using FCM obtain good purity values. Based on these results, this research uses the FCM method to cluster the quality of existing KSM participants to find schools that have good quality education.

49.2 Fuzzy C-Means (FCM) Clustering

Fuzzy c-means (FCM) is a method of clustering data whose membership value determines the existence of each data point in a cluster [15]. The centroid is used to determine the cluster of data. In the initial condition, the cluster center is not accurate and needs to be repaired repeatedly until an accurate centroid is obtained based on the membership value [17, 18]. The output of the FCM is the membership value of each data point and the row of the centroid [19].

Clusters in FCM have a stage that is almost the same as fuzzy logic with the initial stage to determine the value of the membership function of a data [20]. Initial membership value (μ_{ik}) of the FCM is defined in a random manner to determine and update the membership value (μ_{ik}) using Eq. (1).

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}} \tag{49.1}$$

where X_{ij} is the initial data for the clustering, and V_{kj} is the centroid in FCM. Variable i is the number of the data and c is the number of clusters. Variable w is the weight of the data whose default value is 2 [13]. Centroid in FCM can be calculated using Eq. (2).

$$V_{kj} = \frac{\sum_{i=1}^{n}\left((\mu_{ik})^w * X_{ij}\right)}{\sum_{i=1}^{n}(\mu_{ik})^w} \tag{49.2}$$

The FCM method will continue to iterate until it meets two conditions, namely the iteration (t) and the objective function (P_t). The condition for FCM to cease based on the objective function is $P_{t-1} - P_t <$ error (ε) [15]. The objective functions used in the FCM method can be seen in Eq. (3).

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right](\mu_{ik})^w\right) \tag{49.3}$$

49.2.1 FCM Algorithm

The steps to determine clusters using the FCM method [21]:

1. Input the data to be clustered, such as a matrix of size $n \times m$ where the number of data samples (n), attributes per data (m) and ith sample data with $i = 1, 2, \ldots, n$ and j-attribute with $j = 1, 2, \ldots, m$ (X_{ij}).

2. Determine the number of clusters = c, weight = w, maximum iteration = *MaxIter*, error = ε, initial objective function ($P0= 0$), and initial iteration ($t = 1$).
3. Generating initial membership degrees with random numbers (μ_{ik}) with $i = 1$, 2,..., n and $k = 1$, 2,..., c; as an initial partition matrix element (U). Calculate the number of columns from the matrix using Eq. (4).

$$Q_i = \sum_{k=1}^{c} \mu_{ik} \tag{49.4}$$

with $j = 1$, 2,..., n. calculate the random matrix can be seen in Eq. (5).

$$\mu_{ik} = \frac{\mu_{ik}}{Q_i} \tag{49.5}$$

4. Calculate the center of the cluster using Eq. (2).
5. Calculate the objective function using Eq. (3).
6. Calculate changes in the value of membership values using Eq. (1).
7. The FCM process stops on this condition:
   a. $P_{t-1} - P_t <$ error (ε)
   b. *MaxIter* limit reached
8. Define a cluster of data using Euclidean distance.

49.3 Silhouette Evaluation

The silhouette evaluation was proposed by Rousseeuw in 1987 [22]. The method is often used to measure the quality of a cluster. An illustration of the silhouette evaluation can be seen in Fig. 49.1.

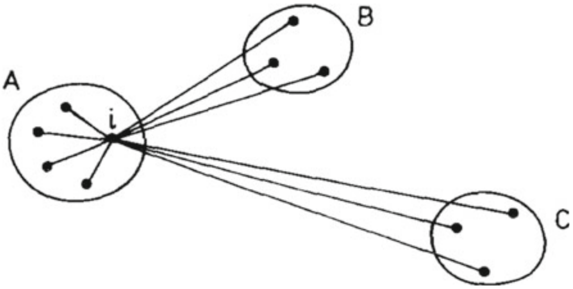

Fig. 49.1 Silhouette evaluation illustration [23]

In Fig. 49.1 can be seen in the data x_i located in Clusters A. Regions B and C are clusters other than A. In the silhouette evaluation, there is also a_i which is the average length of data x_i with data A and b_i is the minimum of average line length between data x_i and data from each other clusters [24]. Based on the illustration in Fig. 49.1, we get the silhouette value of the x_i data [22].

49.4 Result and Discussion

This study uses KSM data in 2019 with a total of 11,457 data with the distribution of data in 6 subjects which is biology, physics, chemistry, mathematics, geography, and economics with the amount of data for each study such as 1116, 1105, 1106, 1125, 1120, and 1117. The data contains the score of the correct answer, the score of the wrong answer, and the score of the wrong answer. Based on these parameters, clustering can occur regarding the quality of education in each school. The score data used is KSM participants consisting of several provinces with each province, and it can be used as a sample of the quality of high school education in every school in the regions in Indonesia. Samples of the distribution of data used can be seen in Fig. 49.2.

In Fig. 49.2, the green lines indicate data with a score of questions that were answered incorrectly. The blue color indicates the score data of the questions that were answered correctly, and the data in orange are the score data of the blank answer. The data will be clustered, including using 3 clusters, 4 clusters, 5 clusters, and 6 clusters. Then, the results of clustering were evaluated using silhouettes. The results used to measure the evaluation of clustering are the value of the silhouette (Si) and also the distribution of data on the silhouette using standard deviation (Std). Table 49.2 shows the results of the clustering process with 4 experiments.

Table 49.1 shows that the highest silhouette values limit to 1 are found in clustering with 3 clusters. The high silhouette value with a low standard deviation represents good clustering results with small data simulations. Figure 49.3 shows the results of the silhouette of each subject area with 3 clusters. The results of silhouette values are more than 0.50 which indicates good structure category according to the Kaufman table. The next process is determining the label of a cluster. The technique that can

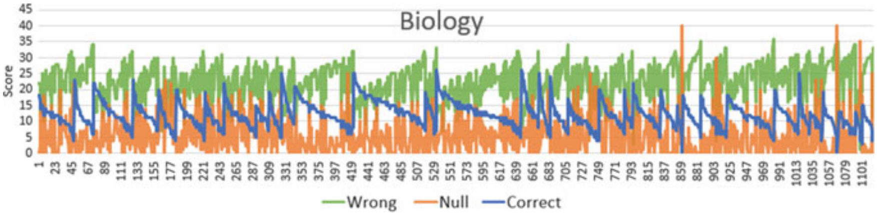

Fig. 49.2 Sample data on KSM scores

Table 49.1 Results of clustering using the FCM method

Subject	3 Clusters		4 Clusters		5 Clusters		6 Clusters	
	Si	Std	Si	Std	Si	Std	Si	Std
Mathematics	0.6499	0.2419	0.5783	0.2497	0.5910	0.2427	0.5642	0.2585
Chemistry	0.6195	0.2667	0.5266	0.2658	0.5385	0.2699	0.5364	0.2847
Geography	0.5220	0.2962	0.5265	0.3099	0.4933	0.3144	0.5049	0.3054
Physics	0.5941	0.2648	0.5539	0.2571	0.5542	0.2654	0.5764	0.2812
Economy	0.5773	0.3253	0.4937	0.3091	0.5161	0.3273	0.5361	0.3196
Biology	0.5663	0.2659	0.5406	0.2671	0.5285	0.2807	0.5143	0.2994

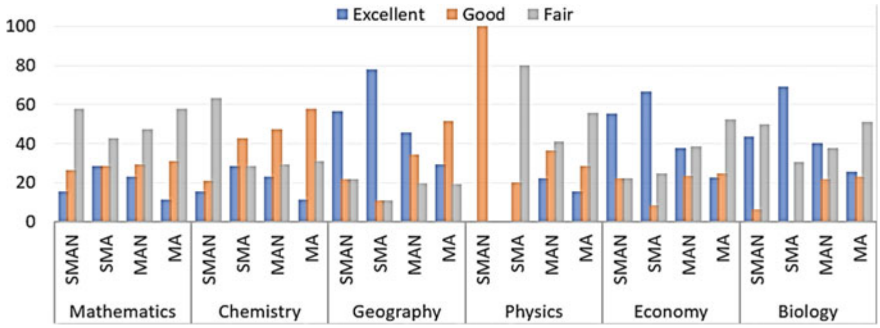

Fig. 49.3 Education quality identification results for each school

be used is to identify data through a cluster center to identify education quality, based on the results obtained by using 3 clusters in Table 49.2.

Table 49.2 carried out the process of defining labels, namely excellent, good, and fair. Labeling is obtained by identifying the cluster center and analyzing the data. The analysis is carried out to determine cluster 1 included in the label excellent, good, or fair. After the labeling process has been completed, next is to identify the distribution of schools in Indonesia to be identified in each subject based on State Senior High School (SMAN), Private Senior High School (SMA), Private Islamic Senior High School (MA), or State Islamic Senior High Schools (MAN) who are superior in terms of quality of education and schools can know the level of success in teaching and educating their students. The results of each Senior High School included in each cluster that has been labeled can be seen in Table 49.3. The table contains the number of SMAN, SMA, MA, and MAN in each existing cluster.

Table 49.3 shows the distribution of education quality data in each high school based on subject areas, specifically mathematics (math), physics (phy), biology (bio), chemistry (chem), geographic (geo), and economy (eco). The results of identifying the quality of each school in Indonesia can be seen in Fig. 49.3.

In Fig. 49.3, it can be seen that in mathematics almost all schools are weak in this subject. In mathematics, the school categorized best in high school with a percentage

Table 49.2 Centroid of clustering result

Subject	Cluster	Data			Label
		Wrong	Null	Correct	
Mathematics	Cluster 1	14.7775	5.7261	4.4964	Good
	Cluster 2	19.3779	0.4471	5.1751	Fair
	Cluster 3	7.8210	14.780	3.0010	Excellent
Chemistry	Cluster 1	22.5352	0.7791	6.6857	Fair
	Cluster 2	16.4525	7.3475	6.2001	Good
	Cluster 3	9.0388	16.7895	4.1717	Excellent
Geography	Cluster 1	15.7173	14.0193	10.2634	Good
	Cluster 2	19.3867	4.5296	16.0837	Fair
	Cluster 3	26.5779	1.3441	12.0780	Excellent
Physics	Cluster 1	13.8807	5.6817	5.4376	Good
	Cluster 2	8.2458	12.5996	4.1546	Excellent
	Cluster 3	18.6352	0.5096	5.8552	Fair
Economy	Cluster 1	19.9990	0.8497	9.1513	Fair
	Cluster 2	12.0580	2.7344	15.2077	Excellent
	Cluster 3	12.1813	8.8399	8.9789	Good
Biology	Cluster 1	27.5889	1.5281	10.8830	Fair
	Cluster 2	15.9453	14.0424	10.0124	Good
	Cluster 3	19.4739	5.3286	15.1975	Excellent

Table 49.3 Senior high school distribution for each cluster

High school	Cluster	Subject					
		Math	Phy	Bio	Chem	Geo	Eco
SMAN	Excellent	3	0	7	3	13	10
	Good	5	10	1	4	5	4
	Fair	11	0	8	12	5	4
SMA	Excellent	2	0	9	2	7	8
	Good	2	1	0	2	1	1
	Fair	3	4	4	3	1	3
MAN	Excellent	76	98	207	81	241	179
	Good	96	161	112	161	182	109
	Fair	156	180	195	198	104	180
MA	Excellent	87	101	145	77	164	140
	Good	239	187	132	215	289	153
	Fair	445	363	292	348	108	324

of students at 30% who have high-quality education. In chemistry, almost all schools are at a good level, but at SMAN, some students are dominantly weak toward this subject and the percentage of students who are not experts in chemistry is around 65%. In geography, SMA is the most upper-quality school with a smart student percentage of around 78%. SMA has sufficient quality because almost all students cannot achieve good quality, while SMAN achieves the best quality acquisition in physics with all students having good quality. In economics and biology, it has a percentage that is almost the same as the best quality achieved by SMAN and SMA. Based on the results that have been shown, it can be concluded that SMA is a school that has good quality among other schools. In the comparison of results between SMAN and SMA, the best results are obtained in SMA except in physics. In the comparison of results between MA and MAN, the best results are obtained on MAN. Based on this description, the school that needs evaluation and treatment regarding educational development is the MA.

49.5 Conclusion

This research was conducted to find out education quality that has been applied in several schools in Indonesia. This research is based on KSM 2019 scores using clustering methods with several experiments. The best results are clustering into 3 clusters. Evaluation of cluster performance uses silhouette (si) and standard deviation (std) on each subject data distribution, where math has a value of si = 0.6499; std = 0.2419, chemistry has a value of si = 0.6195; std = 0.2667, geography has a value of si = 0.5220; std = 0.2962, physics has a value of si = 0.5941; std = 0.2648, economics has a value of si = 0.5774; std = 0.3253, and biology has a value of si = 0.5663; std = 0.2559. The results obtained are high schools which are schools with good quality among other schools. In the comparison of results between SMAN and SMA, the best results are obtained in SMA except in physics. In the comparison of results between MA and MAN, the best results are obtained on MAN. Based on this description, the school that needs evaluation and treatment regarding educational development is the MA.

References

1. Sujarwo, S.: Pendidikan Di Indonesia Memprihatinkan. *WUNY UNY*. **15**(1) (2013)
2. Nelson, R., Dawson, P.: Competition, education and assessment: connecting history with recent scholarship. Assess. Eval. High. Educ. **42**(2), 304–315 (2017)
3. Cantador, I., Conde, J.M.: Effects of competition in education: a case study in an e-learning environment (2010)
4. Munirah, M.: Sistem Pendidikan di Indonesia: antara keinginan dan realita. AULADUNA J. Pendidik. Dasar Islam. **2**(2), 233–245 (2015)
5. Kementrian Agama, R.I.: Petunjuk Teknis Kompetisi Sains Madrasah (KSM) Tahun (2016)

6. Biondi, R.L., Vasconcellos, L., de Menezes-Filho, N.A.: Evaluating the impact of participation in the Brazilian Public School Mathematical Olympiad on math scores in students' standardized tests. J. LACEA Econ. (2012)
7. Idrus, M.: Mutu pendidikan dan pemerataan pendidikan di daerah. *Psikopedagogia J. Bimbing. dan Konseling*. **1** (2012)
8. Park, Y., Yu, J.H., Jo, I.-H.: Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute. Internet High. Educ. **29**, 1–11 (2016)
9. Omran, M.G.H., Engelbrecht, A.P., Salman, A.: An overview of clustering methods. Intell. Data Anal. **11**(6), 583–605 (2007)
10. Santra, A.K, Christy, C.J.: Genetic algorithm and confusion matrix for document clustering. Int. J. Comput. Sci. [Internet]. **9**(1), 322–328 (2012). Available from http://ijcsi.org/papers/IJCSI-9-1-2-322-328.pdf
11. Shikalepo, E.E.: School cluster system for quality education in rural namibian schools. Afr. Educ. Res. J. **6**(2), 48–57 (2018)
12. Najdi, L, Er-raha, B.: Use of unsupervised clustering to characterize graduate students profiles based on educational outcomes. Int. J. Comput. Tech. **3**(2) (2016)
13. Chuang, K.-S., Tzeng, H.-L., Chen, S., Wu, J., Chen, T.-J.: Fuzzy c-means clustering with spatial information for image segmentation. Comput. Med. Imaging Graph. **30**(1), 9–15 (2006)
14. Hathaway, R.J., Bezdek, J.C.: Fuzzy c-means clustering of incomplete data. IEEE Trans. Syst. Man, Cybern. Part B. **31**(5), 735–744 (2001)
15. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: The fuzzy c-means clustering algorithm. Comput. Geosci. **10**(2–3), 191–203 (1984)
16. Dutt, A., Aghabozrgi, S., Ismail, M.A.B., Mahroeian, H.: Clustering algorithms applied in educational data mining. Int. J. Inf. Electron. Eng. **5**(2), 112 (2015)
17. Zhou, K., Yang, S.: Exploring the uniform effect of FCM clustering: a data distribution perspective. Knowledge-Based Syst. [Internet] **96**, 76–83 (2016). Available from http://dx.doi.org/10.1016/j.knosys.2016.01.001
18. Novitasari, D.C.R.: Klasifikasi Sinyal EEG Menggunakan metode fuzzy C-means clustering (FCM) dan adaptive neighborhood modified Backpropagation (ANMBP). J. Mat. MANTIK., pp. 31–36 (2015)
19. Kalavathi, P., Dhavapandiammal, A.: Segmentation of Lung Tumor in CT Scan Images using FA-FCM Algorithms. Res. Gate **18**(5), 74–79 (2016)
20. Zhao, Q., Li, X., Li, Y.: X Zhao (2017) A fuzzy clustering image segmentation algorithm based on hidden Markov random field models and Voronoi tessellation. Patt. Recogn. Lett. **85**, 49–55 (2017)
21. Swindiarto, V.T.P, Sarno, R., Novitasari, D.C.R.: Integration of fuzzy C-means clustering and TOPSIS (FCM-TOPSIS) with silhouette analysis or multi criteria parameter data. In: 2018 International Seminar on Application for Technology of Information and Communication, IEEE, pp. 463–468 (2018)
22. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
23. Dolfe, R., Matinzadeh, K.: Investigating skin cancer with unsupervised learning (2019)
24. Aranganayagi, S., Thangavel, K.: Clustering categorical data using silhouette coefficient as a relocating measure. In: International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), IEEE, pp. 13–17 (2007)