Chapter 27
Identify Elementary Student Distribution Based on *Kompetisi Sains Madrasah* Data Using Probabilistic Distance Clustering

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

Abstract Indonesia is a developing country. The quality of human resources also influences the development of a country. The quality of education is one of the benchmarks for the quality of human resources. The quality of human resources can be seen from the quality of education. Improving the quality of education can be done in various ways. One effort designed to improve the quality of education in Indonesia is the provision of educational competitions in each region in Indonesia, and Kompetisi Sains Madrasah (KSM) is one of the pre-eminent competitions that have been designed. From the KSM Competition, a set of student scores is obtained which is a sample of the quality of education in each province. The number of students in educational institutions, especially elementary schools, is increasing which causes the data of students in the system to improve. The data can be grouped based on ability. Grouping is done using PD clustering. This method is one of the hierarchical grouping methods that have good performance. The cluster of students' abilities is very helpful in finding out educational information in the regions making it easier for parties to do special handling. The results of clustering using PD clustering show that three clusters represent the distribution of student's abilities with a silhouette coefficient of 0.5384 and a standard deviation of 0.3506 for mathematics subjects. Silhouette coefficient is 0.4351 and standard deviation is 0.4688 for science subjects.

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

Maryunah
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_27

27.1 Introduction

The country of Indonesia is one of the countries that is experiencing technological and economic development, or it can be called a developing country [1]. The development of a country is also influenced by the quality of human resources [2]. In the last few years, improvement in the quality of human resources is intensively carried out in all regions in Indonesia, especially in the field of education [3]. The quality of education can be seen from the achievement of national examination scores of all students in Indonesia, which can be seen from the statistics of the average test scores written by the Ministry of Education and Culture in 2019 reaching 51.76% [4]. This number shows that the quality of Indonesian education is still low [5]. From all regions in Indonesia, 95% of regions have education quality below 60% of the specified standard. Including large regions such as West Java (51.19%), Central Java (55.88%), and East Java (54.33%) have a low quality of education [4]. So, it is needed to improve the quality of education to support Indonesia for the better.

Improving the quality of education can be done in various ways [2]. One aim intended to improve the quality of education in Indonesia is to hold educational competitions in every region in Indonesia [6]. Competition can trigger a student's interest in learning and a place for students to compete to improve the quality of education. This competition is in the form of academic and non-academic competitions [7]. One of the academic competitions held is the Madrasah Science Competition (KSM) [8].

KSM is a national-level competition for elementary schools to senior high schools [9]. From the KSM Competition, a set of student scores is obtained which is a sample of the quality of education in each province. The students' abilities can be grouped based on students' scores data. The grouping of students' abilities is beneficial in finding out educational information in the regions making it easier for parties to do special handling. This can be called clustering [10].

Clustering is one of the methods in data mining [11]. Data mining is a process of processing money to obtain, explore, and discover hidden knowledge from a dataset or a huge set of data [12]. Clustering in data mining aims to group data based on the characteristics contained in the data into several clusters [13].

Several experts have done some research on clustering. In a study conducted by Harwati, Ardita, and Febriana in 2014 regarding clustering, student performance using the k-mean method showed quite good results with a standard deviation of 1.79 and a mean of 2.64 [14]. In other research concerning the analysis of students' abilities using clustering techniques conducted by Govindasamy and Velmurugan, the study compared several clustering methods, namely K-means, K-medoids, fuzzy c-means (FCM), and expectation-maximization (EM). FCM and EM have good performance compared to k-means and K-medoids [15]. Other research has been carried out by Tortora for solving clustering problems. In 2013, Tortora applied the PDC to the urban wastewater treatment dataset and showed good results. Tortora also compares PDC with other methods such as K-means, and factorial K-means. The results obtained show that using the k-mean method produces two minima, i.e., 39 and 54% [10]. In

the same case, the application of the PD-clustering method shows that it can achieve a convergent minimum JDF so that it can cluster better than other methods [16]. Therefore, FPDC will be used to group students based on their abilities.

27.2 Probabilistic Distance Clustering (PD-Clustering)

Probabilistic distance clustering (PD clustering) is a non-hierarchical algorithm that defines cluster units based on their chances of belonging to a given cluster [17]. PD clustering was first introduced by Ben-Israel and Lyigun in 2008 [15]. Define with $X = x_{i,j}$ a generic matrix with n units and J variables. K represents the number of clusters that are assumed to be non-empty. The center of the cluster is defined as c_k [18]. PD clustering is based on the principle or model of the relationship between $d(x_i, c_k)$ the distance between x_i data from cluster K and $p(x_i, c_k)$ probability for each point held by a cluster with $k = 1, ..., K$, $i = 1, ..., n$ and $j = 1, ..., J$. The relationship between them is the basic assumption of this method, and the probability is inversely proportional to the distance from the center of the cluster; the product between distance and probability is considered constant depending on $F(x_i)$ [18]. The basic assumptions of PD clustering are expressed by the following equation [16].

$$p_{i,k} d_k(x_i) = F(x_i) \tag{27.1}$$

The number of $F(x_i)$ above is called the joint distance function (JDF). A clustering solution is obtained by identifying centers that minimize JDF.

$$\text{JDF} = \sum_{i=1}^{n} \sum_{k=1}^{K} d_k(x_i) p_{ik}^2, \tag{27.2}$$

where $d_k(x_i)$ and $p_{i,k}$ depend on the center of the cluster. The higher the JDF value, the higher the probability for that point to be one cluster. The details regarding PD clustering are explained in detail by Ben-Israel and Lyigun who suggest the use of p^2 in Eq. (27.2), to refine the problem and confirmed convergence [15]. Other papers also show that the center of the cluster is calculated by

$$c_k = \sum_{i=1,...,n} \left(\frac{u_k(x_i)}{\sum_{j=1,...,n} u_k(x_i)} \right) x_i \tag{27.3}$$

where

$$u_k(x_i) = \frac{p_{ik}^2}{d_k(x_i)} \tag{27.4}$$

Minimizing the JDF value and maximizing the probabilistic of each point belonging to only one cluster means that the JDF value at all k centers is zero and always positive elsewhere [15]. Thus, the center can be said to be the global minimum of JDF. There may be other stationary points because these functions are not convex or quasi-convex, but they are saddle points.

In this study, we consider the form [16]

$$d_k(x_i) = \sum_{j=1}^{1} |x_{i,j} - c_{kj}|, \tag{27.5}$$

where $k = 1, ..., K, i = 1, ..., n$, Until Eq. (27.4) becomes:

$$\text{JDF} = \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} |x_{i,j} - c_{kj}| p_{ik}^2 \tag{27.6}$$

From Eq. (27.7), the final solution $\widehat{\text{JDF}}$ is obtained by minimizing the quantity JDF to be

$$\widehat{\text{JDF}} = \min_{C} \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} |x_{i,j} - c_{kj}| p_{ik}^2 \tag{27.7}$$

On the condition that $\sum_{i=1}^{n} \sum_{k=1}^{K} p_{ik}^2 \leq n$ corresponds to Ben-Israel and Lyigun, where c_k are the center of the generic cluster and $0 \leq p_{ik} \leq 1$.

The problem of PD clustering can be solved using an iterative algorithm whose convergence is shown in Lyigun (2007) [18]. Each unit is assigned to cluster k based on the highest probability calculated posterior using the following formula [19]

$$p_{ik} = \frac{\prod_{m \neq k}^{K} d_m(x_i)}{\sum_{l=1}^{K} \prod_{m \neq l}^{K} d_m(x_i)}, \tag{27.8}$$

where $k = 1, ..., K$. Keep in mind that p_{ik} understands every necessary condition as a probability. Besides, no assumptions were made concerning the distribution of this function. This shows that p_{ik} can only be calculated if given x_i and for $\forall c_k$ [20].

27.3 Silhouette Coefficient

The silhouette coefficient is an evaluation method to test the accuracy of a cluster that has been formed from the clustering process. The silhouette coefficient value is defined as in Eq. (27.9)

$$s(i) = \frac{(b(i) - a(i))}{\max a(i), b(i)} \tag{27.9}$$

The results of the silhouette coefficients are interpreted using the Kaufmann and Rousseeuw guidelines for silhouette coefficients.

27.4 Result and Discussion

In this study, identification of the distribution of elementary school students from the data of students who joined KSM to identify the quality of education in an area is based on the value obtained from KSM 2019. In 2019, KSM will split the competition into two subjects, such as mathematics and science. This study uses three parameters obtained from each competition. These parameters are the score of the correct answer, a score of the wrong answer, and the score of the blank answer. The data used in this study were the values of participants who participated in the KSM 2019 competition in each district, totaling 2385 students from 34 provinces with 1189 data in mathematics and 1197 data in science subjects. The data will be clustered to find out the quality of education from each region. This study uses data of students who participated in the KSM primary school level or equivalent. The data consists of the score of correct answers, the score of wrong answers, and the score of blank answers. The data is processed to find a cluster of students. The data is illustrated in Fig. 27.1.

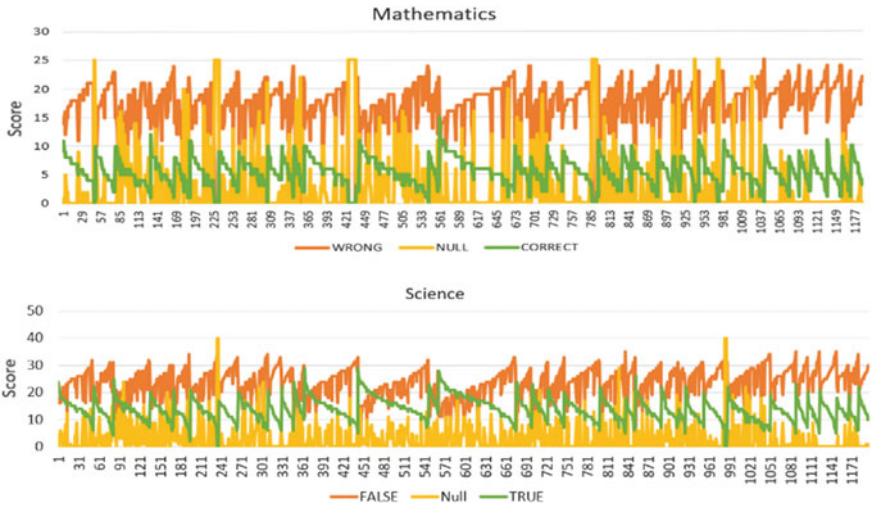

Fig. 27.1 Graph values KSM involving students for science

Figure 27.1 shows that the score chart of students participating in KSM. The orange line represents the score of the wrong answer from the student, yellow represents the blank score of the answer from the student, and the green represents the score of the correct answer from the student. The number of students who answered blanks in mathematics is more numerous than science. The average student answers more wrong than right. This data is used for clustering students who participate in KSM using PD clustering. Initialize some variables needed in PD clustering, such as determining the number of clusters. In this study, the determination of the number of clusters was carried out by trial. The cluster number experiment is intended to see the optimal number of clusters in KSM data grouping. Tests performed four times on each subject. The trial results are evaluated using the silhouette coefficient (Si) and standard deviation (Std) to see the data distribution shown in Table 27.1.

Based on Table 27.1, it can be seen that the optimal number of clusters is three clusters because in the trial of three clusters the largest silhouette value is obtained and the smallest standard deviation compared to the other clusters. A large silhouette coefficient or close to one indicates that each data can be adequately grouped based on its cluster and a small standard deviation means a small data deviation. Figure 27.2 shows the silhouette results of each subject with three clusters.

Figure 27.2 shows the overlapping of the silhouette results, which blends some data that is not in the appropriate cluster. The results of the silhouette of the mathematics subjects were seen overlapping in cluster 1 by 75 data from 1188 data and in

Table 27.1 Clustering results using PD clustering

Number of clusters	Mathematic		Science	
	Si	Std	Si	Std
3	0.5384	0.3506	0.4351	0.4688
4	0.5158	0.4283	0.362	0.4872
5	0.493	0.4333	0.3726	0.529
6	0.3126	0.6402	0.256	0.5348

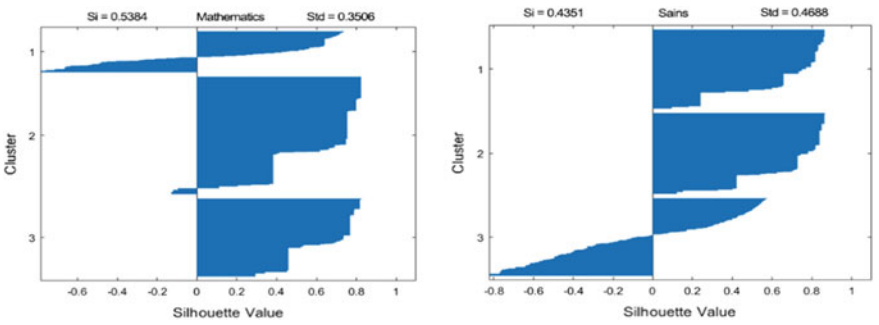

Fig. 27.2 Silhouette results in 3 clusters

cluster 2 by 29 data from 1188 data. In the graphs of science, subjects have 205 overlapping data in cluster 3 and no overlapping in other clusters. Although the results of clustering using PD-clustering are still experiencing overlapping, the results of this study can already be said to be good with silhouette values close to 1. In this research, students are clustered into three clusters of excellent, good, and fair.

In some cases, as shown in Fig. 27.3, some local minimums are obtained. Figure 27.3 shows that the convergence of JDF scores in mathematics subjects starts from the third iteration onwards, whereas in science subjects starting from the second iteration onwards. The JDF value obtained indicates that the cluster center k is said to be the global minimum of JDF. Based on the results obtained using the three clusters in Table 27.2, the cluster center obtained from each cluster, such as cluster 1, cluster 2, and cluster 3 is shown in Table 27.2. The results of identifying the quality of each school in Indonesia are shown in Fig. 27.4.

Figure 27.4 shows that Madrasah Ibtida'iyah (MI) is superior to Elementary Schools (SD) and Madrasah Ibtida'iyah Negeri (MIN) with a percentage of 52% including students with good quality. In science, the average distribution of MI and MIN students has the same quality between excellent, good, and fair while for elementary students 60% of the 9 students participating in KSM. The number of elementary school students participating in KSM is only 2 students, and all of them are clustered into the fair. Based on the results that have been shown, it can be concluded that MI is

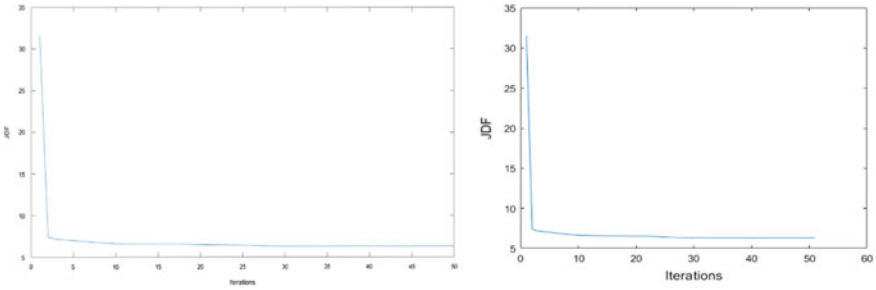

Fig. 27.3 JDF value for each iteration of the PD clustering algorithm

Table 27.2 Cluster center

Subjects	Number of clusters	Data		
		Wrong	Null	Correct
Mathematic	1	4.1594	11.025	9.8153
	2	6	19	2.6985
	3	8	17	4.8282
Science	1	15.99767	23.99824	0.004086
	2	12	28	1.860064
	3	15	19	6.0000000009

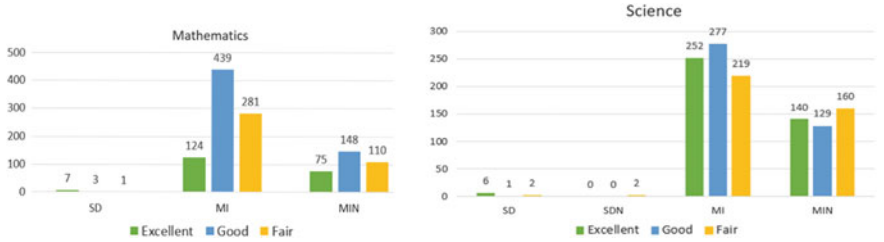

Fig. 27.4 Education quality identification results for each school

an elementary school that has good quality compared to other primary schools. Based on this description, the schools that need evaluation and care regarding improving education are SD and SDN.

27.5 Conclusion

Clustering using the PD clustering method in KSM obtains a maximum number of clusters of 3 clusters because it has the best silhouette coefficient compared to the number of other clusters. These 3 clusters are the first cluster representing fair students, the second cluster represents good students, and the last cluster represents excellent students. Based on Table 27.2, the silhouette coefficient of mathematical subjects produced in this research is intermediate, which means that the cluster structure formed is good while the science subjects have a good cluster structure with a silhouette value of less than 0.5. Based on Fig. 27.4, it can be concluded that MI is an elementary school that has good quality compared to other primary schools. Based on this description, the schools that need evaluation and care regarding improving education are SD and SDN.

References

1. Sutrisno, N.: Pemajuan Kepentingan Negara-Negara Berkembang Dalam SIstem WTO. The Institute for Migrant Rights Press, Indonesia
2. Achola, P.P.: Challenges of primary education in developing countries, 1st edn. Ashgate, New York, NY 10017, USA
3. Kemendikbud. Perbaikan Kualitas Manusia Melalui Pendidikan Dasar dan Menengah. Jakarta
4. Kemendikbud. Laporan Nilai Ujian Nasional Indonesia- Pusat Penilaian Pendidikan [Internet]. Kementrian Pendidik. dan Kebud. (2019). https://puspendik.kemdikbud.go.id/hasil-un/
5. Montoya, S.: Quality data to ensure a quality education for every child (2017)
6. Purba, E.: Menuju Indonesia Baru. Pertama. Guepedia, Jakarta
7. Muhaimin, Manajemen Pendidikan: Aplikasi Dalam Penyusunan Rencana Pengembangan Sekolah/Madrasah. Kelima. Kencana Publisher, Jakarta
8. Basori, R., Arif, Gagasan, F.M.: Ucapan, dan Tindakan dalam Mencerahkan Pendidikan Islam dan Kerukunan Umat. LKis, Yogyakarta

9. Nofrion, Komunikasi Pendidikan: Penerapan Teori dan Konsep Komunikasi dalam Pembelajaran. Pertama. Kencana Perdana Media Group, Jakarta
10. Prasetyo, E.: Data Mining, Mengelola Data Menjadi Informasi Menggunakan Matlab. ANDI Yogyakarta, Yogyakarta
11. Swindiarto, V.T.P.: Integration of fuzzy C-means clustering and TOPSIS (FCM-TOPSIS) with Silhouette analysis for multi criteria parameter data. In: 2018 International Seminar on Application for Technology of Information and Communication, pp. 463–468 (2018)
12. Prasetya, E.: Data mining Mengolah Data Menjadi Informasi Menggunakan MATLAB. Andi, Yogyakarta
13. Sindhu, R., Nandal, R., Dhamija, P., Sehrawat, H.A.: Review on K-means algorithm and its' different distance matrics. Int. J. Eng. Technol. [Internet] **9**, 1423–1430 (2017). https://doi.org/10.21817/ijet/2017/v9i2/170902227
14. Permata, A., Ayu, F.: Mapping student' s performance based on data mining approach (A case study). Ital. Oral Surg. [Internet] **3**, 173–177 (2015). http://dx.doi.org/10.1016/j.aaspro.2015.01.034
15. Govindasamy, K.: Analysis of student academic performance using clustering techniques **119**(15), 309–323 (2018)
16. Tortora, C., Gettler, M., Marino, M., Palumbo, F.: Factor probabilistic distance clustering (FPDC): a new clustering method. Adv. Data Anal. Classif. (2015)
17. Iyigun, C., Ben-israel, A., Iyigun, C.: Sciences I. Sciences : probabilistic distance clustering adjusted for cluster size (September 2008), 603–621 (2015)
18. Lyigun, C.: Probabilistic distance clustering (2007)
19. Van Den, P.D.: Algorithms from and for nature and life classification and data analysis
20. Rachev, S., Klebanov, L., Stoyanov, S., Fabozzi, F.: A new dimension reduction method: factor discriminant k-means. J. Classif. **2**(28), 210–226 (2013)