Journal of Engineering and Applied Sciences 12 (13): 3367-3371, 2017
ISSN: 1816-949X

Profiling the Educational Development in Indonesia Using a Fuzzy Clustering Approach

Lizda Iswari
Department of Informatics, Universitas Islam Indonesia,
Jalan Kaliurang Km. 14.4, 55501 Yogyakarta, Indonesia

Abstract: Education is a key indicator to measure the development of a nation and the quality of its human resources. Indonesia has a vast number of school components which become a challenge for the government on how to define the appropriate policies to improve the equal educational opportunities throughout the country. The level of educational development can be quantified by education index involving a number of parameters. The most common analysis of these parameters is primarily based on the statistical distribution of data. However, this current analysis has some limitations to extract the keys information that are necessary for the policy makers to recognize the educational needs that could be differ in each region. This study aims to define groups of province that have similar profiles of educational development. Research was done by clustering the education index data set which consist of 33 provinces and 15 parameters. The clustering method was based on a fuzzy clustering algorithm which was able to provide the uniformity of each generated cluster. For determining the optimum number of clusters it was based on a cluster validity measurement, i.e., Xie and Beni Index. Meanwhile for interpreting the clustering results it was based on student t-test since it involved the data comparison to the national education standards. This research showed that fuzzy clustering can be utilized to identify the cluster of provinces based on their level of achievement in education and provide some insights that might be used by the policy makers in determining the educational development priorities.

Key words: Clustering, education, fuzzy, student, profile

INTRODUCTION

Education is a key indicator to measure a nation’s development and its quality of human resources. Similarly, Indonesia is aware that education must be put as the main priority of development. Indonesia has a vast number of school components where currently there are 50 million students, 2.6 million teachers and 250.000 schools spread over 33 provinces. This becomes a challenge for the government on how to define the appropriate policies that improves the equal educational opportunities throughout the country.

To quantify the achievement of education development programs at the national and provincial levels the government of Indonesia through ministry of education and culture has defined some parameters of education index. The most common analysis of these parameters primarily is based on the basic statistical information by taking some average of data and categorize them into five categories based on the crisp partition to represent the education level for each administrative areas. Unfortunately, this approach is considered to have some limitations including the incapability to show which variables are dominant and trigger the education index into a higher or lower values, incompetence to recognize the real problems that exist in the education domain which may tailor an inability to identify the educational needs that could be differ in each regions.

One approach to overcome these problems can be done by cluster analysis since it does not rely on common assumptions to conventional statistical methods and is useful in situations where little prior knowledge exists (Chattopadhyay *et al*., 2011). Cluster analysis is aimed at determining the group of data based on similar characteristics. The further development of this analysis also considers the level of membership which include the fuzzy sets as the weighting basis for data grouping named as fuzzy clustering. This method is an extension of crisp partitioning (k-means algorithm) that enable the data objects belong to several clusters simultaneously with different degree of membership (Ferraro and Giordani, 2015). In fuzzy clustering, data have a tendency to be a member of a cluster with the highest degree of membership (Chattopadhyay *et al*., 2011). Fuzzy clustering has been widely used for analysing a data set since it can provide a smooth and effective results and can improve the uniformity of each of generated clusters (Shihab, 2000).

Table 1: Parameters of the national education standard

Mission	Parameters	Abbrv.	Unit	EL	MH	HS	Explanation
K1	The ratio of students per school	RSS	Student	168	288	384	EL 6 class, MH 9 class, HS 12 class
	The ratio of students per class	RSC	Student	28	32	32	Gov. Rule Nr.15/2010
	The ratio class per classroom	RCC	Class	1	1	1	Ideal
	The percentage of library	LIB	Percentage	100	100	100	Ideal
	The percentage of laboratory	LAB	Percentage	-	100	100	Ideal
K2	The level of school services	SER	Student	46	89	78	National standard 11/12
	The affordable area	AFR	Student	181	376	576	National standard 11/12
K3	The percentage of decent teachers	PDT	Percentage	100	100	100	Ideal
	Number of graduates	NGR	Percentage	100	100	100	National standard 11/12
	Number of students repeating grades	SRG	Percentage	0	0	0	Ideal
	Number of dropouts	NDO	Percentage	0	0	0	Ideal
	The percentage of good classroom	PGC	Percentage	100	100	100	Ideal
K4	Gender differences of GER	GDF	Percentage	0	0	0	Ideal
	Gender parity index of GER	GPI	Index	1	1	1	Ideal
K5	Gross enrollment rate (GER)	GER	Percentage	100	100	100	Ideal

Previously, there were several researches that dealt with the analysis of Indonesia education index as found by Widayati *et al.* (2010) which mapped the priority of education development planning using WebSOM (Self Organizing Maps). She classified 17 districts of a county into five levels of education coverage using six indicators of education index. Unfortunately, she did not describe any reasons why the data set was divided into five classes. A larger use of parameters can be found on Kapita and Irawan that utilized the Kohonen SOM to cluster 50 elementary schools from three counties based on school's self-assessment. This assessment consisted of 31 indicators. This research was able to show some school clusters based on their education quality toward the national education standards.

This study aimed to define the groups of province that have similar profiles of educational development. Research was done by analysing the education index data set which consist of 33 provinces based on fuzzy clustering approach. There were 15 parameters considered in line with the national education missions. The result of clustering was analyzed and compared to the national education quality standard to identify the cluster of provinces based on their level of achievement in the field of education.

The standard of educational quality in Indonesia: Indonesia defines three types of Education Level, i.e., Elementary (EL), Middle High (MH) and Senior High (HS). To quantify the achievement of educational development program the government of Indonesia has defined a number of parameters that are in line with the national education mission. This mission has been divided into 5 groups, i.e., K1 as the availability of education services, K2 as the affordability of education services, K3 as the quality of educational services, K4 as the equivalency to obtain the educational services and K_5 as the certainty of obtaining educational services.

In total there are 20 parameters, however, only 15 parameters are available as shown in Table 1. Every parameters is provided with the national education quality standards as a basis for evaluating the level of educational achievement for each administrative area.

MATERIALS AND METHODS

This research was conducted in four steps as follows.

Data pre-processing: As seen in Table 1 the education data set has a mix of attributes, i.e., different unit of data and different range of values which could impact the results of clustering. Hence, a data normalization is required to enhance the clustering quality (Thangavel and Visalakshi, 2009). This research applied a min-max normalization which performed a linear transformation on the original data into a specified range of values, i.e., all values to be transformed between 0-100. The min-max normalization was based on Eq. 1 as follows:

$$v' = \frac{v\text{-}\min_a}{(\max_a\text{-}\min_a)}(\text{new }\max_a\text{-new }\min_a)+\text{new }\min_a \tag{1}$$

Where:

v = A value to be normalized

Min_a and $\max_a$ = The minimum and the maximum values of the data set

New_min_a and new_max_a = The desired range of data

Data clustering based on fuzzy clustering: Fuzzy clustering in this research was done by using the open source software R version 3.2.4 with the function of FKM (Fuzzy K-Means) algorithm that belongs to fclust package. The FKM function was originally defined by

Bezdek in 1974 which assigned objects to clusters according to membership degree in [0, 1] (Ferraro and Giordani, 2015). The functions works as Eq. 2 and 3:

$$\min_{UH} J_{FKM} = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^{m} d^2\left(x_i, h_g\right) = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^{m}\left|x_i - h_g\right|^2 \quad (2)$$

$$\text{s.t.} = u_{ig} \in \left[0, 1\right], \sum_{g=1}^{k} u_{ig} = 1 \quad (3)$$

Where:

$X = [x_{ij}]$ = Data matrix of order (n×t)
$U = [u_{ig}]$ = Membership degree matrix of order (n×k)
$H = [h_{ig}]$ = Prototype matrix of order (k×t)
m (>1) = Parameter of fuzziness
n = Number of objects
t = Number of variables
k = Number of clusters

Selecting the cluster number: To select the optimum number of cluster, Xie and Beni index was chosen as the basis of measurement (Chan *et al*., 2007) states XB Index has a high accuracy and reliability to provide the optimum number of cluster in fuzzy clustering. The cluster validity index by Xie and Beni (1991) are defined as Eq. 4:

$$XB(k) = \frac{\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^{m} d^2\left(h_i, h_g\right)}{n\min_{g.g'(g\neq g)} d^2\left(h_g, h_{g'}\right)} \quad (4)$$

Clustering evaluation: The clustering results were analysed based on the statistical inference, i.e., hypothesis testing with a single sample's mean. In more specific terms a two-tailed test where the region of rejection is on both sides of the sampling distribution. The calculation of sampling distribution was based on student t-test, since the number of samples of each distribution or cluster was <30 (n<30). The student t-test was based on Eq. 5 (Box *et al*., 2005):

$$t = \frac{x-\mu_0}{s/\sqrt{n}} \quad (5)$$

Where:

x = The sample mean
μ_0 = The specified data mean
S = The sample standard deviation
n = Sample size

RESULTS AND DISCUSSION

The FKM algorithm was applied to cluster the educational profiles for middle high. Firstly, data was normalized into the range of 0-100. Secondly, data was clustered started from 2-10 clusters by using the FKM algorithm. In this research, we set the parameter of fuzziness (m) equal to two this was related to Klawonn and Keller which mentioned the most optimum and widely used in many research of fuzziness parameter is two. Furthermore, the convergence criterion was set equal to 1e-9 and the maximum number of iterations was equal to 1e+6. The result of clustering is shown in Table 2 whereas it showed some additional information of the number of iterations to achieve data convergence the time of computation and the Xie and Beni Index to show the cluster validity index.

Table 2: Cluster validity index

No. of cluster	No. of iteration	Computation time	Xie and Beni index
2	57	0.16	0.8803878
3	177	0.72	8.141787e+17
4	210	1.15	8.299882e+17
5	409	2.75	6.592994e+22
6	200	1.56	1.695927e+21
7	197	1.82	2.185605e+20
8	222	2.29	1.170153e+20
9	256	3.03	8.667919e+19
10	335	4.35	1.241568e+22

Prior to the data analysis, the selection of the optimum number of cluster was done based on the Xie and Beni (XB) Index. Chen *et al*. (2007) stated that the optimal cluster number on XB index was determined by the first gap, i.e., the first minimum value on a series of trials. As shown in Table 2 the first gap was found on cluster six as the first value that decline from nine trials of clustering.

Afterward, the FCM algorithm was applied to cluster the data set into six clusters and retrieved some essential information such as the membership degree of all data objects and the cluster center of all parameters. Although, data has been clustered into six clusters we have found there were four clusters with similar values both in the membership degrees and the cluster centers. Hence, we divided the data set into three clusters: C1, C2 and C3. The assignment of data into a specific cluster was done by choosing the highest degree among these three clusters. Overall, there were 20 provinces labelled as C_1, six provinces labelled as C2 and seven provinces labelled as C3.

The next step was the cluster analysis by taking the cluster center of all parameters and categorized their level of appropriateness to the national standard values. To measure the level of appropriateness we used the student t-test as mentioned on Eq. 5 including the cluster centers as the sample mean (x) the national standard values as the specified data mean (μ_0) the standard deviations of each clusters (s) and the number of object of each cluster (n). The result of t-score is shown in Table 3. The t-test is a

Table 3: Cluster center, t-score and parameter categorization results

Param	Cluster center of C1	Cluster center of C2	Cluster center of C3	t-score of C1	t-score of C2	t-score of C3	Category of C1	Category of C2	Category of C3
RSS	25.469	64.428	48.167	-10.918	1.643	-0.569	NG	G	G
RSC	42.067	24.193	29.616	5.016	1.280	2.508	NG	G	NG
RCC	65.411	26.480	29.580	7.575	-1.064	-0.678	NG	G	G
LIB	30.992	70.510	38.048	-54.352	-7.176	-40.876	NG	NG	NG
LAB	28.292	57.846	38.924	-17.500	-2.115	-6.595	NG	G	NG
SER	46.078	44.008	57.788	-1.453	-1.077	0.432	G	G	G
AFR	27.346	69.253	57.721	-9.186	1.147	-0.024	NG	G	G
PDT	70.397	91.111	77.157	-6.744	-5.113	-6.937	NG	NG	NG
NGR	92.960	97.814	97.645	-1.444	-3.380	-2.266	G	NG	G
SRG	42.264	11.046	19.822	9.016	3.463	5.846	NG	NG	NG
NDO	29.129	11.485	16.924	7.051	4.570	5.015	NG	NG	NG
PGC	53.036	77.755	63.818	-10.788	-9.899	-6.930	NG	NG	NG
GDF	54.699	59.506	67.693	-0.058	0.908	1.057	G	G	G
GPI	42.840	38.015	30.592	-0.003	-1.215	-1.007	G	G	G
GER	41.952	65.433	52.523	-1.774	1.424	0.241	G	G	G

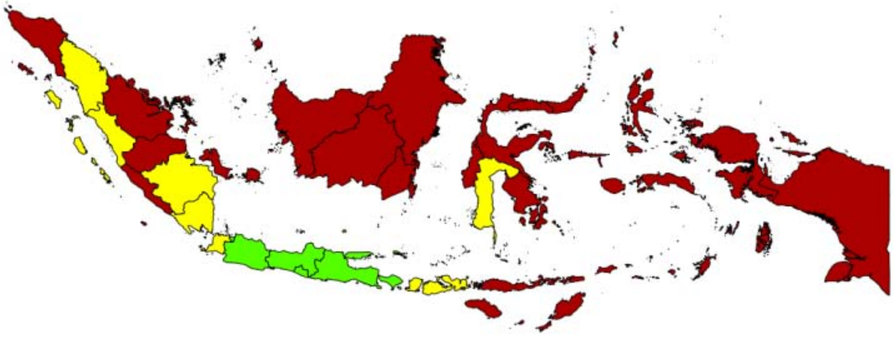

Fig. 1: The distribution of cluster in spatial perspective

comparative test to determine if two sets of data are significantly different to each other, i.e., a comparison between cluster center to the national standard values. The test on the null hypothesis will be rejected if the cluster center is significantly higher or lower than the national standard values.

The result of t-test was used to define the region of rejection on both sides of the sampling distribution, i.e., the rejection limits on two-tails. To define these limits, we needed a t-normal distribution table which was dependent upon the number of samples and the significance level. By using the significance level of 0.05 (99.5%) the rejection limits of each clusters is as:

- Cluster C1 with 20 samples: ±2.086
- Cluster C2 with 6 samples: ±2.447
- Cluster C3 with 7 samples: ±2.365

Based on the limits above, we categorized the cluster center toward the standard values, i.e, if the t-score was lower or higher than the rejection limits it would be labelled as Not Good (NG) and conversely, if the parameters fall in between the limits would be labelled as Good (G). The G label indicates that parameters in compliance with the national education standards while the NG label indicates that these parameters do not meet the standards that have been set. The result of parameters categorization for each clusters was shown in Table 3.

As seen in Table 3, there were four out of 15 parameters (27%) categorized as G in all clusters, i.e., SER, GDF, GPI and GER. Meanwhile, the higher number was found on NG since there were five parameters (33%) categorized as NG, i.e., LIB, PDT, SRG, NDO and PGC. These parameters should have more attention from the government since none of the provinces are able to achieve the standard of educational quality.

Furthermore, the distribution of parameter categorization of each clusters may also be used to determined the educational development priorities. Among these three clusters, we defined that.

C1 is the group of provinces with the lowest level of educational achievement since 67% of parameters (10 of 15) were categorized as not good. These included RSS, RSC, RCC, LIB, LAB, AFR, PDT, SRG, NDO and PGC. This cluster also had the largest member with 20 provinces dominated by provinces in Kalimantan Island and eastern part of Indonesia. These provinces were drawn in dark orange color as seen on Fig. 1.

C2 is the group of provinces with the best level of educational achievement for being dominated by the Good category. Nevertheless, it still had 40% (6 of 15) of parameters below the national standard, i.e., LIB, PDT, NGR, SRG, NDO and PGC. This cluster had six members, all of which are located in Java Island as shown in green color in Fig. 1.

C3 is the group of provinces with the moderate level of educational achievement since the number of good parameters is almost equal to the not good. This cluster had 47% (7 of 15) of not good and had seven provinces located spread over in Sumatera Island and central part of Indonesia. This cluster was drawn in yellow color as seen in Fig. 1.

CONCLUSION

Based on the research results, there are some conclusions that can be drawn: this research has applied the Fuzzy K-Means (FKM) to cluster all provinces in Indonesia based on the similar similar charateristics of education index profiles. The most optimum number was three clusters whereas the educational needs on each cluster can be identified based on a comparison between the level of achievement of parameters and the national standard of education.

The evaluation of clustering results was based on student t-test able to categorize the parameters into two classes, i.e., good category to indicate the parameters in compliance with the national standards and not good category to indicate parameters that did not meet the national standards.

REFERENCES

Box, G.E., W.G. Hunter and J.S. Hunter, 2005. Statistics for Experimenters: Design, Innovation and Discovery. 2nd Edn., John Wiley and Sons, New York.

Chattopadhyay, S., D.K. Pratihar and D.S.C. Sarkar, 2011. A comparative study of fuzzy c-means algorithm and entropy-based fuzzy clustering algorithms. Computi. Inf., 30: 701-720.

Chen, D., X. Li and D.W. Cui, 2007. An adaptive cluster validity index for the fuzzy c-means. Intl. J. Comput. Sci. Network Secur., 7: 146-156.

Ferraro, M.B. and P. Giordani, 2015. A toolbox for fuzzy clustering using the R programming language. Fuzzy Sets Syst., 279: 1-16.

Shihab, A.I., 2000. Fuzzy clustering algorithms and their application to medical image analysis. Ph.D Thesis, University of London, London, England.

Thangavel, K. and N.K. Visalakshi, 2009. Impact of normalization in distributed K-means clustering. Int. J. Soft Comput., 4: 168-172.

Widayati, N., M. Hariadi and S. Mardi, 2010. Mapping the planning priority based on websom multidimensional data mining. Sepuluh Nopember Institute of Technology, Surabaya, Indonesia.

Xie, X.L. and G. Beni, 1991. A validity Measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell., 13: 841-847.