The Clustering of High Schools Based on National and School Examinations

A Case Study at Daerah Istimewa Yogyakarta Province

Paulina H. Prima Rosa, Ridowati Gunawan
Department of Informatics
Sanata Dharma University
Yogyakarta, Indonesia
rosa@usd.ac.id; rido@usd.ac.id

Ignatius Aris Dwiatmoko
Department of Mathematics
Sanata Dharma University
Yogyakarta, Indonesia
aris.dwiatmoko@usd.ac.id

***Abstract*— The purpose of Indonesian National Examination for High School Students is to measure and assess students' knowledge and competence in particular subjects. The result is also going to be used as one of consideration for mapping Indonesia's national education quality. Aside from National Examination (NE), each school also conducts School Examination (SE). Both examinations are supposed to represent quality of education since the examinations measure the competence of the same students. However, the results of both examinations are not always linear [1]. In fact, the need of NE in Indonesian education is still being pro and cons among society.**

In order to identify whether NE and SE could be used to represent the quality of educations in Daerah Istimewa Yogyakarta Province, this paper describes the analysis of NE and SE score by performing data mining technique using Fuzzy C-Means clustering algorithm towards NE score and SE score independently. Furthermore, the clusters were analyzed using univariate Anova, Spearman correlation, and crosstabulation. The data used in this research are NE and SE scores of Natural Science Department and Social Science Department of all high schools in Daerah Istimewa Yogyakarta Province from academic year 2011/2012 to 2014/2015.

The results of cluster analysis are three different clusters of NE in Natural Science Department, three different clusters of NE in Social Science Department, three different clusters of SE in Natural Science Department, and three different clusters of SE in Social Science Department for each year. The clusters are significantly separated. There is an opposite direction relationship between clusters of NE and SE. The relationship is weak which means there is no guarantee that a school which belongs to cluster-i of NE will be in the cluster-i of SE. Both for NE and SE memberships, only few members migrated from one cluster to another across years. The number of schools having the same cluster of NE and SE in each department varies from year to year, but generally less than 22%. The migrations of NE and SE cluster members from higher cluster to lower one and vice versa also vary.

In addition, there is inconsistency clustering based on NE and SE. Since SE is not standardized and indeed is a formative test, there might be subjective aspects involved in grading the students. Therefore, if the government intends to map Indonesia's national education quality, national examination is more suitable than school examination for this purpose.

Keywords*— *knowledge discovery in databases; Indonesian National Examination; Clustering; fuzzy c-means; Anova; Spearman correlation; crosstab

I. INTRODCUTION

Several efforts in the area of curriculum, quality and professionalism of teachers, infrastructure, and evaluation systems have been done by Indonesian government to improve the quality of education in Indonesia. In term of evaluation, Indonesian government has established a government law (PP no. 19/2005 article 63) containing several types of evaluation in education, namely evaluation by teachers, evaluation by schools, and evaluation by government as well.

Evaluation by teachers is a formative evaluation during the learning process in the whole semesters. Evaluation by schools are usually performed at the end of each semester and usually referred as School Examination (SE). Evaluation by government, which is usually referred as National Examination (NE), is performed at the end of the period of study in particular levels (elementary school, secondary school, and high school). Exam questions in NE are standardized by government, while exam questions in SE are composed by each school. In addition, each school has an authority to determine minimum completeness criteria for their students to pass the exams.

Based on PP no. 19/2005 article 63, the purpose of Indonesian National Examination for High School Students is to measure and assess students' knowledge and competence in particular subjects. The result is also going to be used as one of consideration for mapping Indonesia's national education quality.

In the year of 2010, the result of NE was the only factor to determine whether students should pass from a particular level of education or not. In this period, there were 267 schools whose students 100% failed in NE (Kompas, 28/4/2010). During

academic year 2011/2012 until 2014/2015, both NE and SE scores contributed to the Final Score (FS) that was used to determine students' graduation from High School.

Both NE and SE are supposed to represent quality of education since the examinations measure the competence of the same students. However, the authors found that the results of both examinations are not always linear [1]. In fact, the need of NE in Indonesian education is still being pro and cons among society. Therefore, the authors conduct a further research to identify whether NE and SE could be used to represent the quality of educations by taking a case study at Daerah Istimewa Yogyakarta Province.

Several studies about NE in particular places in Indonesia have been performed as can be found in [1], [2], [3], and [4]. Educational data mining using clustering methods have also been used by [5], [6],[7],[8],[9] and [10] for mapping the quality of schools based on the results of NE.

Other algorithms that are also frequently used is Fuzzy Clustering Means (FCM). FCM algorithm was first proposed by Ruspini [11] and later updated by Dunn [12] and Bezdek [13]. FCM clustering is a technique in which the existence of each data point in a cluster is determined by the degree of membership. The purpose of the Fuzzy C-Mean algorithm is to find the centroid by minimizing the objective function. FCM has been used for geo statistical data analysis problems in [13] while Lu et.al. in [14] used FCM for meteorological data.

In the domain of education in Indonesia, research using Fuzzy C-Means algorithm had been done by Luthfi [15] who clustered lecturers' teaching performance. Fuzzy C-Means for distribution of lecture participants is carried out by Susanto [16]. Karti in [17] used C-Means algorithm and Fuzzy C-Means clustering to cluster cities in East Java Province based on the similarity of high schools and vocational schools education indicators, namely the Net Enrolment Ratio (NER) and the Gross Enrolment Ratio (GER).

However, the researchers have not found researches that study NE in comparison with SE. The map of school quality based on NE and SE, the characteristics of each clusters as well as the consistency of each cluster across years have not been found yet. As the following part of the first research in [1], in this paper the researchers describe the analysis of NE score and SE score by performing data mining technique using Fuzzy C-Means clustering algorithm towards NE score and SE score independently. The clusters were then analyzed using univariate ANOVA, Spearman correlation, and crosstab. Univariate Anova was used to validate whether the clusters are significantly separated. Spearman correlation was used to identify the relationship between NE cluster memberships and SE cluster memberships. In addition, Spearman correlation was also used to identify the membership of each cluster across several years as well. Crosstabulation was used to find the percentage of schools whose memberships in NE clusters are the same with the memberships in SE clusterss.

The data used in this research are NE and SE scores of all high schools in Yogyakarta Province from academic year 2011/2012 to 2014/2015.

The result of this research hopefully could be used to identify whether NE and SE could be used to represent the quality of educations in Daerah Istimewa Yogyakarta Province.

II. Research Methodology

This research uses Knowledge Discovery in Databases [18] process. The implementation of the methodology is as follows:

1. Selection process as the process to select target data. The data available from the website of research and development division of The Indonesian Ministry of Education and Culture. The data can be downloaded in the form of spreadheet. It contains the result of SE for primary schools, secondary schools, and high schools in all provinces in Indonesia from year 2011 until 2014 and the result of NE as well as FS from year 2010 until 2014. For this research, the data selected as target data are high schools' SE and NE from year 2011 until 2014. The data from year 2010 is not included due to the fact that it does not contain SE since NE is the only factor to determine the students graduation during year 2010. The data of NE and SE are limited only for Natural Science Department and Social Science Department of all high schools (state and private schools) in Daerah Istimewa Yogyakarta Province. The province was selected as the case study since the province is the residence of the researchers so that it is more feasible to realize the plan to conduct in depth analysis in the future. Language Department was excluded as target data since only several schools have language department. The data of Natural Science Department contains fifteen columns as described in Table I, while the data of Social Science Department contains almost similar colums as described in Table II.
2. Preprocessing process to clean noise in the data. Theoretically, in this step the data should be cleaned from noises. However, the data that were downloaded from official website of The Indonesian Ministry of Education and Culture did not contain any noise. Therefore, there was no data cleaning performed. However, it was identified that the number of schools from year 2011 to 2014 were not the same due to the facts that there were several new schools in particular years or there were no students in several schools in particular years. These facts would not affect analysis, so that there is no need to perform data cleaning.
3. Transformation process to convert data from the official website into particular form that is ready to be mined. Since the data were clustered independently for each year and each type of exam (NE/SE) for each department, there was no transformation needed to convert data as described in Table I and Table II. Therefore, there were 16 dataset in the form of spreadsheet as described in Table III.

TABLE I. DATA DESCRIPTION OF NATURAL SCIENCE DEPARTMENT

No	Field
1	Number
2	School code
3	School name
4	School status (state/private)
5	Number of students joining examination
6	Number of students failed
7	The percentage of students failed
8	Mean score of Indonesian language test
9	Mean score English test
10	Mean score Mathematics test
11	Mean score Physics test
12	Mean score Chemical test
13	Mean score Biology test
14	Total of mean score of six courses being tested in Natural Science Department
15	The rank of school

TABLE II. DATA DESCRIPTION OF SOCIAL SCIENCE DEPARTMENT

No	Field
1	Number
2	School code
3	School name
4	School status (state/private)
5	Number of students joining examination
6	Number of students failed
7	The percentage of students failed
8	Mean score of Indonesian language test
9	Mean score English test
10	Mean score Mathematics test
11	Mean score Ecoomics test
12	Mean score Sociology test
13	Mean score Geography test
14	Total of mean score of six courses being tested in Social Science Department
15	The rank of school

4. Data mining process by performing data mining technique using Fuzzy C-Means clustering algorithm towards NE score and SE score independently. The algorithm was implemented using Java programming language. The program read 16 datasets that have been prepared in step 3, one file at a time. In each running, the program was executed using the following parameters: number of clusters = 3, weighting exponent = 2, maximum number of iterations = 100, and error tolerance = 0.001. Since the research would identify the consistencies of clustering across years, in each running the program read each dataset in Table III independently. The results of the program are member of each clusters along with the centroid of each clusters.
5. Evaluation to analyze the clusters formed in step 4 were then performed by using several analytical tools as follow:
   a. Univariate Anova to validate that the clusters are significantly separated.
   b. Spearman correlation to identify the relationship between NE cluster memberships and SE cluster memberships. In addition, Spearman correlation is also used to identify the membership of each cluster across several years as well.
   c. Crosstabulation to find the percentage of schools whose memberships in NE cluster as well as SE cluster remain the same.
6. Interpretation was done to describe the meaning of the results of the above analytical tools so that it can be easily understood by common people.

III. RESULTS AND DISCUSSIONS

A. Fuzzy C-Means Clustering

In this research, clustering was performed to cluster schools based on NE and SE scores. The clustering will classify schools such that schools with similarity will be in the same cluster. The clusters will have high internal homogeneity and high external heterogeneity as well. To be more effective, cluster analysis was preceded by outlier detection using histogram. Based on the histograms, there was no outlier founded in the data of NE and SE scores for both natural science and social science departments.

By using Fuzzy C-Means, schools were classified into 3 groups according to the categorization of schools determined by The Ministry of Education and Culture, namely group A, B, and C that represent schools with high score category, middle score category, and low score category respectively. Table IV to Table VII describe the centroid of each cluster for each year. Cluster 1 represents schools with the highest score, cluster 2 represents schools with middle score, while cluster 3 represents the lowest one. From the table, it can be identified that the number of schools in each cluster varies. Several schools migrate from cluster 1 to 2, 2 to 3, and vice versa. The migration of schools can be identified from the membership of each cluster.

TABLE III. DATASET OF NE AND SE

No	Name of dataset	Type of Exam	Department	Year
1	NE_ND_11	National Exam	Natural Science	2011
2	NE_ND_12	National Exam	Natural Science	2012
3	NE_ND_13	National Exam	Natural Science	2013
4	NE_ND_14	National Exam	Natural Science	2014
5	NE_SD_11	National Exam	Social Science	2011
6	NE_SD_12	National Exam	Social Science	2012
7	NE_SD_13	National Exam	Social Science	2013
8	NE_SD_14	National Exam	Social Science	2014
9	SE_ND_11	School Exam	Natural Science	2011
10	SE_ND_12	School Exam	Natural Science	2012
11	SE_ND_13	School Exam	Natural Science	2013
12	SE_ND_14	School Exam	Natural Science	2014
13	SE_SD_11	School Exam	Social Science	2011
14	SE_SD_12	School Exam	Social Science	2012
15	SE_SD_13	School Exam	Social Science	2013
16	SE_SD_14	School Exam	Social Science	2014

TABLE IV. NE CLUSTER CENTROID OF NATURAL SCIENCE DEPARTMENT

	Cluster		
	1	2	3
Year 2011	7.72	6.61	5.12
Number of schools	52	50	32
Year 2012	7.58	6.40	5.00
Number of schools	54	53	28
Year 2013	7.72	6.30	5.25
Number of schools	32	59	46
Year 2014	7.20	5.76	4.58
Number of schools	34	51	56

TABLE V. NE CLUSTER CENTROID OF SOCIAL SCIENCE DEPARTMENT

	Cluster		
	1	2	3
Year 2011	7.61	6.32	4.85
Number of schools	65	65	34
Year 2012	7.54	6.21	4.90
Number of schools	52	64	47
Year 2013	7.76	6.19	5.03
Number of schools	31	61	70
Year 2014	7.45	5.82	4.57
Number of schools	35	55	69

TABLE VI. SE CLUSTER CENTROID OF NATURAL SCIENCE DEPARTMENT

	Cluster		
	1	2	3
Year 2011	8.86	8.36	7.89
Number of schools	45	62	27
Year 2012	9.03	8.45	7.99
Number of schools	38	49	48
Year 2013	9.16	8.58	8.11
Number of schools	28	56	53
Year 2014	9.17	8.55	8.12
Number of schools	19	59	63

TABLE VII. SE CLUSTER CENTROID OF SOCIAL SCIENCE DEPARTMENT

	Cluster		
	1	2	3
Year 2011	8.78	8.35	7.80
Number of schools	43	86	35
Year 2012	8.91	8.45	8.03
Number of schools	47	66	50
Year 2013	9.04	8.54	8.13
Number of schools	37	73	52
Year 2014	9.21	8.56	8.18
Number of schools	29	65	65

B. Univariat Anova

To make sure that the clustering based on NE and SE result on clusters which is significantly independent, Univariat Anova was performed on each academic year by checking whether there are differences of NE and SE mean scores between clusters.

Significant differences indicate that clustering is able to significantly differentiate schools into 3 groups.

Based on Anova tables, it can be concluded that the clustering of schools based on the mean score of NE and SE for all academic years have resulted on three significantly different clusters. Cluster 1, 2, 3 consecutively are schools with the highest mean score of NE or SE (first order), second order, and third order. As an illustration, Table VIII is an example of Anova table for NE score in year 2011. Significance value 0.000 (which is less than $\alpha = 0.05$) shows that the clustering has resulted on significantly different clusters. All Anova tables for NE score and SE score in year 2011 to year 2014 has significance value 0.000.

C. Spearman Correlation

The using of Spearman Correlation is intended to identify the relationship between membership of NE and SE clusters. High correlation between the two shows there is a consistency of clusters based on NE and SE. Ideally, a school that is categorized in cluster 1 of NE will be categorized as cluster 1 of SE as well.

The result of Spearman correlation of NE and SE membership for both natural science and social science department are described in Table IX and X. Based on those tables, negative correlation indicates that there is an opposite relationship between clusters based on NE and SE. The correlation is weak (the absolute values is less than 0.5) which means there is no guarantee that a school which belongs to cluster-i of NE will be in the cluster-i of SE. Negative and weak correlation also shows that several schools which are belong to high NE score cluster will tend to be members of lower SE score cluster. Few schools in particular SE cluster migrate into lower NE cluster.

Using Spearman correlation it can also be identified the cluster membership correlation across years. Higher correlation indicates that schools will tend to be in the same cluster. For NE, the membership correlation which is between 0.581 to 0.855 (for natural science department) and 0.670 to 0.868 (for social science department) shows that only few members migrated from one cluster to another across years. The conclusion also applies for SE as represented by the membership correlation which is between 0.541 to 0.758 (for natural science department) and 0.503 to 0.800 (for social science department).

D. Crosstabulation

The using of crosstab is to support the Spearman correlation. From crosstab it can be identified the percentage of schools whose memberships in NE clusters are the same with the membership in SE clusters. It can also be counted the percentage of schools whose memberships change. Table XI shows an example of crosstabulation between NE and SE clusters in year 2011. Crosstab for other years are not presented in this paper

TABLE VIII. ANOVA TABLE OF NE IN YEAR 2011

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	130.916	2	65.458	453.030	.000
Within Groups	18.828	131	.144		
Total	149.845	133			

TABLE IX. SPEARMAN CORRELATION BETWEEN NE AND SE OF NATURAL SCIENCE DEPARTMENT

	Cluster_SE_11	Cluster_SE_12	Cluster_SE_13	Cluster_SE_14
Cluster_NE_11	-0.307			
Cluster_NE_12		-0.261		
Cluster_NE_13			-0.400	
Cluster_NE_14				-0.376

TABLE X. SPEARMAN CORRELATION BETWEEN NE AND SE OF SOCIAL SCIENCE DEPARTMENT

	Cluster_SE_11	Cluster_SE_12	Cluster_SE_13	Cluster_SE_14
Cluster_NE_11	-0.217			
Cluster_NE_12		-0.275		
Cluster_NE_13			-0.412	
Cluster_NE_14				-0.454

TABLE XI. CROSSTABULATION OF NE CLUSTERS AND SE CLUSTERS OF NATURAL SCIENCE DEPARTMENT IN YEAR 2011

			Cluster_SE11			Total
			1	2	3	
Cluster_NE11	**1**	**Count**	13	20	19	52
		% of Total	9.7%	14.9%	14.2%	38.8%
	2	**Count**	15	29	6	50
		% of Total	11.2%	21.6%	4.5%	37.3%
	3	**Count**	17	13	2	32
		% of Total	12.7%	9.7%	1.5%	23.9%
Total		**Count**	45	62	27	134
		% of Total	33.6%	46.3%	20.1%	100.0%

Based on crosstabulations, it can be recognized several facts as follows:

1. The number of schools having the same cluster of NE and SE in each department varies from year to year, but generally less than 22%.
2. The migrations of NE and SE cluster members from higher cluster to lower one and vice versa also vary.

The above facts emphasize that there is inconsistency clustering based on NE and SE. As elaborated in [1], the clustering of NE will tend to end up in clusters with wide interval of NE score, while clustering of SE will end up in clusters with narrow interval and high score of SE.

Since SE is not standardized, there might be subjective aspects involved in grading the students. Since SE indeed is a formative test, schools would try their best to help students passing minimum completeness criteria and finally combine with NE score that will contribute to the final score of student's graduation. Interview with several educational practitioners revealed these facts.

Therefore, if the government intends to map Indonesia's national education quality, national examination is more suitable form of test for this purpose.

IV. Conclusions

The results of cluster analysis are three different clusters of NE in Natural Science Department, three different clusters of NE in Social Science Department, three different clusters of SE in Natural Science Department, and three different clusters of SE in Social Science Department for each year. The clusters are significantly separated. There is an opposite direction relationship between clusters of NE and SE. The relationship is weak which means there is no guarantee that a school which belongs to cluster-i of NE will be in the cluster-i of SE. Therefore, it should be carefully examined which is actually represent the quality of education, either NE score or SE score.

Both for NE and SE memberships in all departments, only few members migrated from one cluster to another across years. The number of schools having the same cluster of NE and SE in each department varies from year to year, but generally less than 22%. The migrations of NE and SE cluster members from higher cluster to lower one and vice versa also vary.

Future works will be performed to analyze clusters toward each subject in NE to identify the possible strengths and weaknesses of each school. Analysis of new data from the year 2015 in which NE are not used as component to determine student's graduation will also be performed to study the effect of the new policy towards students and school achievement.

Acknowledgment

The authors thank to the following:

1. Directorate of Higher Education, Ministry of Education and Culture for funding the research.
2. Maria Riana Megawati and Renia Octaviani for gathering data and preparing the software.

References

[1] A. Dwiatmoko, P. Rosa, and R. Gunawan, "Analisis Statistis Data Nilai Ujian Nasional dan Nilai Ujian Sekolah Menengah Atas di Propinsi Daerah Istimewa Yogyakarta," *J. Ilm. Media Tek.*, vol. 14, no. 2, pp. 89–95, 2015.

[2] Rasto, Nahadi, and S. A. Soesilawaty, "Analisis Peta Kompetensi Hasil Ujian Nasional dan Model Pengembangan Mutu Sekolah SMA di Jawa Barat (Survey di Kabupaten Garut dan Kabupaten Tasikmalaya)," *Artik. Has. Penelit.*, vol. 2011, pp. 1–17, 2011.

[3] A. Jaedun and E. Listiyani, "Pemetaan Mutu Pendidikan di SMA Kabupaten Cilacap dan Banyumas Berdasarkan Analisis Hasil Ujian Nasional," in *Prosiding Seminar Nasional*, 2012.

[4] Zulkardi, "Model Peningkatan Mutu Pendidikan SMA di Kota Prabumulih, Kabupaten Ogan Ilir, dan Kabupaten Ogan Komering Ilir," in *Prosiding Seminar Nasional Hasil Penelitian Universitas Sriwijaya*, 2011.

[5] R. . Khairunnisa and H. Fajriya, "Pengelompokan Sekolah Dasar Berdasarkan Rata-Rata Nilai Ujian Akhir Sekolah Berstandar Nasional Di Setiap SD/MI di Kota Yogyakarta Dengan Menggunakan Metode Hierrarkhi Clustering," in *Naskah Seminar Kerja Praktek, FMIPA, UII Yogyakarta*, 2013.

[6] Pakpahan and M. Juni, "Pengelompokan Sekolah Menengah Pertama Berdasarkan Rata-Rata Nilai Ujian Akhir Nasional Di Kota Binjai dengan Analisis Hierrarkhi Clustering," in *Naskah Skripsi Universitas Negeri Medan*, 2013.

[7] Prihatiningtyas, "Analisis Hasil Ujian Nasional Tingkat SMA di Kabupaten Banyumas Menggunakan Analisis Cluster dan Biplot," in *Skripsi Program Studi Statistika, Universitas Diponegoro, Semarang,* 2011.

[8] Habibi and A. Ahmad, "Pemetaan UN Propinsi di Indonesia Melalui Analisis Klaster Penggabungan Pautan Tunggal," in *Thesis JBPTITBPP, ITB Bandung, Indonesia*, 2009.

[9] Fadhli, "Analisis Kluster Untuk Pemetaan Mutu Pendidikan di Aceh," in *Thesis Jurusan Matematika Universitas Gadjah Mada Yogyakarta*, 2011.

[10] Sjafrudin and Asep, "Analisis Hasil Ujian Nasional Madrasah Tsanawiyah Tahun 2008," 2009.

[11] E. H. Ruspini, "A new approach to clustering," *Inf. Control*, vol. 15, no. 1, pp. 22–32, 1969.

[12] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, 1973.

[13] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy *c*-means clustering algorithm," *Comput. Geosci.*, vol. 10, no. 2–3, pp. 191–203, 1984.

[14] Y. Lu, T. Ma, C. Yin, X. Xie, W. Tian, and S. Zhong, "Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data," *Int. J. Database Theory Appl.*, vol. 6, no. 6, pp. 1–18, 2013.

[15] E. T. Luthfi, "Fuzzy C-Means Untuk Clustering Data (Studi Kasus: Performance Mengajar Dosen)," *Semin. Nas. Teknol. 2007 (SNT 2007)*, vol. 2007, no. November, pp. 1–7, 2007.

[16] S. Susanto and Ernawati, "Pembagian Kelas Peserta Kuliah Berdasarkan Fuzzy Clustering dan Partition Coefficient and Exponential Separation (PCAES) Index," *J. Teknol. Ind.*, vol. X1 (2), no. APRIL 2007, pp. 143–154, 2007.

[17] H. S. Karti, "Pengelompokan Kabupaten / Kota di Provinsi SMA / SMK / MA dengan Metode C-Means dan Fuzzy C-Means," *J. Sains dan Seni Pomits*, vol. 2, no. 2, p. D–288, 2013.

[18] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, First. Boston, MA, USA: Addison-Wesley Longman Publishing, 2005.