ISSN 2621 – 086X

Classification of Human Development Index Using *K*-means

Retno Tri Vulandari[1] , Sri Siswanti[2], Andriani Kusumaningrum K[3], and Kumaratih Sandradewi[4]
[1, 2] Informatics Engineering, STMIK Sinar Nusantara, Surakarta, Indonesia
[3, 4] Accounting Computer, STMIK Sinar Nusantara, Surakarta, Indonesia

vulandari.sinus@gmail.com

Abstract. Human development progress in Central Java. It is characterized by a continued rise in the human development index (HDI) of Central Java. HDI is an important indicator for measuring success in the effort to build the quality of human life. HDI explains how residents can access the development results in obtaining a long and healthy life, knowledge, education, decent standard of living and so on. HDI is affected by four factors, namely life expectancy, expected years of schooling, means years of schooling, and expenditure per capita. Currently the Central bureau of statistics do grouping HDI, using calculation formula then known how the value HDI each regency or city in Central Java. In this research we classified the regency or city in Central Java based on the HDI be high, middle, and under estimate area. We used cluster analysis. Cluster analysis is a multivariate technique which has the main purpose to classify objects based on their characteristics. Cluster analysis classifies the object, so that each object that has similar characteristics to be clumped into a single cluster (group). One of the cluster analysis method is *k*-means. The result of this research, there are three groups, high estimate area, middle estimate area, and under estimate area. The first group or the under estimate area contained 12 regencies, namely Cilacap, Purbalingga, Purworejo, Wonosobo, Grobogan, Blora, Rembang, Pati, Jepara, Demak, Pekalongan, and Brebes. The second group or the middle estimate area contained 8 regencies, namely Banjarnegara, Kebumen, Magelang, Temanggung, Wonogiri, Batang, Pemalang, and Tegal. The third group or the high estimate area contained 11 regencies, namely Banyumas, Kudus, Boyolali, Klaten, Sukoharjo, Karanganyar, Sragen, Semarang, Kendal, Surakarta, and Salatiga.

Keywords : cluster analysis, *k*-means, the human development index.

1. Introduction

The success of human development can be shown from human development index (HDI). HDI is a measurement comparison of a long and healthy life, knowledge, education, and decent standard of living for all countries or regions throughout the world. HDI were used to classify whether a country is a developed country or underdeveloped countries and also to measure the impact of the economic wisdom of the quality of life. Since 2001 the central government gave autonomy to the local government authority. Each local government is drawing up development planning and the financial budget of his country, not only to increase economic growth, but also improve the welfare of society through human development which includes the sectors of education, health care, and other policies that directly improve the quality of life [1]. Thus, it can improve the well-being of the community through improved human development indicators are reflected in HDI [2].

Changes of HDI in Indonesia are indicated by the composite index but not indicated which variables are dominant against the high/low HDI ranks. Grouping of regencies/cities in Central Java have to be done as a planning and evaluation targets of government programs to increase the number of human development. Politically the area aims to divide the territories in several groups with characteristics that have a high degree of similarity in every group and has the difference between groups [3].

Research advance of the HDI, commenting about how to obtain the HDI is affected by life expectancy (LE), expected years of schooling (EYS), means years of schooling (MYS), and expenditure per capita (EPC) with regression analysis method. The results obtained the four factors affecting the magnitude of HDI and the results of the grouping obtained based on high low HDI value predicted [4].

In this study, grouping will be discussed based on the HDI of the four factors by looking at the value of the similarity of characteristics formed. Analysis of clustering (cluster analysis) is one of the dual variables analysis (multivariate analysis) that is used to collect objects that have the semblance of a characteristic in one group. Cluster analysis consist of group hierarchical structure (hierarchical clustering) and nonhierarchical clustering. One of the nonhierarchical clustering is *k*-means [5]. Then the mapping will be done with the inverse distance weighted interpolation-based.

2. Clustering Method

Clustering can be considered the most important unsupervised learning problem, so as every other problem of this kind, it deals with finding a structure in a collection of unlabelled data. A loose definition of clustering could be “the process of organizing objects into groups whose regency are similar in some way” [6]. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

For clustering algorithm to be advantageous and beneficial some of the conditions need to be satisfied. In this case we easily identify the 4 clusters into which the data can be divided. The similarity criterion is distance: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance). This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures [7]. The following Table 1 gives literature review on *k*-means and its modifications.

Table 1. *K*-means and Its Enhancement

No	Author(s)	Description
1	Pethalakshmi, 2012	Modification in *k*-means clustering algorithm through affinity measure to increase the cluster uniqueness [8]
2	Goryawala, 2012	Analyzed on 3-D liver segmentation using combined approach of *k*-means and segmentation algorithm [9]
3	Siddiqui, 2012	The optimized *k*-means (OKM) *cluste*ring algorithm using "An Image Segmentation" with the capability of avoiding the Dead Centre and Trapped Centre at local minims [10]
4	Vlase, 2012	An improvement of *k*-means clustering algorithm using various patent metadata [11]
5	Li, 2012	Evaluates clustering in image indexing using *k*-means clustering algorithm and genetic algorithm [12]

Pethalakshmi [8] said that for real time large database it was difficult to predict the number of cluster and initial seeds accurately. In order to overcome this drawback, new algorithms were proposed. They were unique clustering through affinity measure (UCAM) and it worked without fixing initial seeds, number of resultant cluster to be obtained and unique clustering was obtained with the help of affinity measures.

Goryawala [9] have done 3-D segmentation that based on coupling a modified *k*-means segmentation method with a special localized contouring algorithm. In the segmentation process, five separate regions were identified on the computerized tomography image frames. The merit of the proposed method layed in its potential to provide fast and accurate liver segmentation and 3-D rendering as well as in delineating tumor regions, all with minimal user interaction. Leveraging of multi-core platforms is shown to speed up the processing of medical images considerably, making this method more suitable in clinical settings. Experiments were performed to assess the effect of parallelization using up to 442 slices. Empirical results, using a single workstation, showed a reduction in processing time from 4.5 h to almost 1 h for a 78% gain. Most important was the accuracy achieved in estimating the volumes of the liver and tumor region(s), yielding an average error of less than 2% in volume estimation over volumes generated on the basis of the current manually guided segmentation processes. Results were assessed using the analysis of variance statistical analysis.

Siddiqui [10] despite the fact that the previous improvements of the conventional *k*-means algorithm. It considers the modification on the hard countyhip concept as employed by the conventional *k*-means algorithm. As the process of a pixel is assigned to its associate cluster, if the pixel has equal distance to two or more adjacent cluster centres, the pixel will be assigned to the cluster with null (e. g., no regency) or to the cluster with a lower fitness value. The qualitative and quantitative analyses had been performed to investigate the robustness of the proposed algorithm. It is concluded the new approach was effective to avoid dead centre and trapped centre at local minima which led to produce better and more homogenous segmented images. The OKM algorithm had been introduced as the modified version of the conventional *k*-means algorithm. The proposed OKM algorithm fully concentrated on differentiating between the dead centers and zero variance clusters (e. g., cluster with similar intensity pixels). The pixel with the same distance to two or more adjacent clusters was initially assigned to the dead centre and in later iteration, it was assigned to the cluster with lower variance cluster, if no dead centre could be found. The proposed OKM algorithm was capable to produce good quality segmentation with more homogeneous regions of interest. The convincing results reveal that the proposed OKM algorithm proffers excellent consistency in its performance and can be used in different electronic products as an image segmentation tool.

Vlase, et al. [11] said that by controlling the influence of applicants, clusters that contain only the relevant patent applicants, and not all their patents, could be obtained. The higher the influence of applicants in clustering was, the more patents of the same applicant appeared in the same cluster, decreasing the content similarity of patent obtained exclusively by classical clustering. This also applicable to the other metadata fields, where the title or the inventor could be mentioned. The application area was for patent databases, but the idea could be generalized and also investigated in other areas. Those aiming to improve the quality of clustering by emphasizing the importance of various patent metadata, that could be achieved by computing different weights for different patent metadata attributes, which were considered to be valuable information.

Li [12] described that the processing speed of *k*-means was the fastest, but not on the largest data set. *K*-means was the best choice in the conditions that emphasized on time but not asked for much on the accuracy. A simple genetic algorithm has better clustering effects for the problem with little amount of categories and data. But when the parameter and amount of categories increased, the effects became bad. It did not apply for complicated clustering problems. The performance of new methods improved from 30% to 50% comparing to simple *k*-means method. The time consumed by new method was between *k*-means and genetic algorithm. Because of restrictions of genetic algorithm, the scale of data processed could not be too large and the calculations were increased greatly. The algorithm must be optimized or use parallel method to improve efficiency.

3. Results and Discussion

The steps of data classification is as follows.

1. Determine the center of the cluster randomly, in this study the initial cluster centers was taken 3 regency, which showed under estimate, middle estimate, and high estimate area, as in Table 2.

Table 2. The Initial Cluster Centers

	Regency	LE	MYS	EYS	EPC
C1	Brebes	68.20	5.88	11.34	8898
C2	Wonogiri	75.86	6.39	12.42	8417
C3	Klaten	76.55	8.16	12.84	11178

Based on Table 2, Brebes represents under estimate area, Wonogiri represents middle estimate area, and Klaten represents high estimate area. It based on the report of Central Bureau of Statistics (BPS) [13].

2. Calculate the distance of each existing data against any cluster center. Suppose the first data to calculate the distance:

Distance the first data to the first cluster center:

$$d_{11} = \sqrt{(73 - 68.2)^2 + (6.58 - 5.88)^2} = 453.02$$

Distance the first data to the second cluster center:

$$d_{12} = \sqrt{(73 - 75.86)^2 + (6.58 - 6.39)^2} = 934.0$$

Distance the first data to the third cluster center:

$$d_{13} = \sqrt{(73 - 76.55)^2 + (6.58 - 8.16)^2} = 1827$$

All calculation of the distance can be seen in Table 3.

Table 3. The Results of The Cluster Distance

Distance	d1	d2	d3
Cilacap	453.03	934.00	1827.00
Banyumas	1206.01	1687.00	1074.01
Purbalingga	40.28	521.01	2240.00
Banjarnegara	968.01	487.01	3248.00
Kebumen	890.01	409.01	3170.00
Purworejo	407.05	888.00	1873.00
Wonosobo	838.00	1319.01	1442.01
Magelang	716.02	235.01	2996.00
Boyolali	2908.01	3389	628.00
Klaten	2280.02	2761.00	0
Sukoharjo	1518.03	1999.00	762.00
Wonogiri	481.06	0	2761.00
Karanganyar	1588.03	2069.00	692.00
Sragen	2536.01	3017	256.01
Grobogan	559.04	1040.00	1721.00
Blora	199.08	282.01	2479.00
Rembang	224.08	705.00	2056.00
Pati	482.06	963.00	1798.00
Kudus	1305.03	1786.00	975.00
Jepara	606.05	1087	1674.00
Demak	220.12	701.00	2060.00
Semarang	1880.01	2361	400.00
Temanggung	529.05	48.01	2809.00
Kendal	1521.01	2002.00	759.00
Batang	654.03	173.01	2934.00
Pekalongan	310.04	791.00	1970.00
Pemalang	1721.01	1240.00	4001.00
Tegal	531.01	50.25	2811.00
Brebes	0	481.06	2280.02
Surakarta	4706.01	5187.00	2426.00
Kota Salatiga	5702.01	6183.00	3422.00

Table 3 shows the similarity of distance in each regency/city. For example the similarity distance of Surakarta with cluster center 1 is 4706.01, cluster center 2 is 5187, and cluster 3 is 2426.

3. Data will become a member of a cluster that has the smallest distance from the cluster center. For example, for first data, the smallest distance is obtained on the first cluster,

so the first data will be a member of the first cluster. Position the cluster more information can be seen in Table 3. Table 3 shows the position of group on the first iteration. On the first iteration, Cilacap is on the first group, Banyumas is on the third group, Purbalingga is on the first group, etc.

4. Calculate the new cluster center. For the first cluster, there are 12 so that life expectancy (LE) can be stated by,

$$c_{11} = \frac{(73\ + 72.81 + \ldots\ + 68.2)}{12} = 73.43$$

Means years of schooling (MYS),

$$c_{12} = \frac{(6.58\ + 6.85 + \ \ldots\ + 5.88)}{12} = 6.69$$

Expected years of schooling (EYS),

$$c_{13} = \frac{(12.28\ + 11.78 + \ \ldots\ + 11.34)}{12} = 12.04$$

Expenditure per capita (EPC),

$$c_{14} = \frac{(9351\ + 8938 + \ \ldots\ + 8898)}{12} = 9226.33$$

For the second cluster, there are 8 so that life expectancy (LE),

$$c_{21} = \frac{(73.59 + 72.77 + \ \ldots\ + 70.9)}{8} = 73.61$$

Means years of schooling (MYS),

$$c_{22} = \frac{(6.17\ + 7.04 + \ \ldots\ + 6.3)}{8} = 6.51$$

Expected years of schooling (EYS),

$$c_{23} = \frac{(11.39\ + 12.49 + \ \ldots\ + 12)}{8} = 11.91$$

Expenditure per capita (EPC),

$$c_{24} = \frac{(7930\ + 8008 + \ \ldots\ + 8367)}{8} = 8086.75$$

For the third cluster, there are 11 so that life expectancy (LE),

$$c_{31} = \frac{(73\ .12 + 75.63 + \ \ldots\ + 76.83)}{11} = 75.93$$

Means years of schooling (MYS),

$$c_{32} = \frac{(7.31\ + 7.1 + \ \ldots\ + 9.81)}{11} = 8.03$$

Expected years of schooling (EYS),

$$c_{33} = \frac{(12.57\ + 12.13 + \ \ldots\ + 14.97)}{11} = 13.08$$

Expenditure per capita (EPC),

$$c_{34} = \frac{(10104 + 11806 + \dots + 14600)}{11} = 11366.18$$

5. Repeat step 2 until the position of the data has not been changed. In this study on the 2nd iteration no change occurred. The first group or the cluster contained 12 regencies, i.e. Cilacap, Purbalingga, Purworejo, Wonosobo, Grobogan, Rembang, Blora, Pati, Jepara, Demak, Pekalongan, and Brebes. The second cluster contained 8 regencies i.e. Temanggung, Kebumen, Banjarnegara, Tegal, Wonogiri, Magelang, Batang, and Pemalang. The third cluster contained 11 regencies i.e. Kudus, Banyumas, Boyolali, Klaten, Sukoharjo, Karanganyar, Sragen, Semarang, Surakarta, Kendal, and Salatiga.

Figure 1 is an explanation of a visual grouping of HDI, there are 12 under estimate area shows with blue, there are 8 middle estimate area shows with green, and there are 11 high estimate area shows with purple.

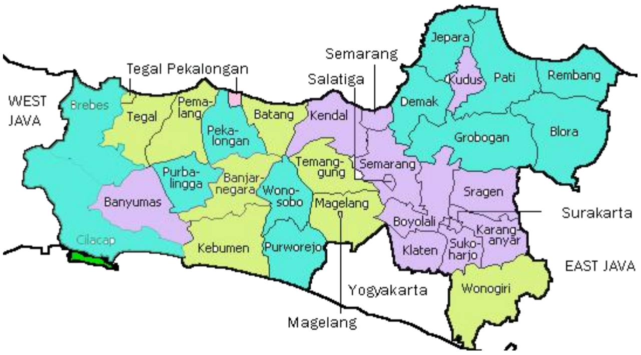

Figure 1. HDI in Central Java 2015

4. Conclusion

The *k*-means algorithm can be used to classify the HDI by looking at the characteristics of the constituent factors of similarity. The *k*-means algorithm can save a complicated calculation process that becomes the reference of calculation of previous HDI. The *k*-means algorithm can distinguish each group for sure based on the similarity of characteristics, not just from the previous intervals the benchmark value. The first cluster contained 12 regencies, i.e. Cilacap, Purbalingga, Purworejo, Wonosobo, Grobogan, Rembang, Blora, Pati, Jepara, Demak, Pekalongan, and Brebes. The second cluster contained 8 regencies, i.e. Kebumen, Tegal, Banjarnegara, Wonogiri, Magelang, Batang, Temanggung, and Pemalang. The third cluster contained 11 regencies i.e. Kudus, Banyumas, Boyolali, Klaten, Sukoharjo, Karanganyar, Sragen, Semarang, Surakarta, Kendal, and Salatiga. This research is the classification of the human development index in Central Java that is affected by four factors, namely life expectancy, hope of the old school, old school average, and the average real per capita expenditure by the method of *k*-means. For further research can also be developed with other cluster method, such as *k*-nearest neighbors (KNN), mixture modelling, self-organising map (SOM), and rock algorithm.

Acknowledgements

The authors gratefully acknowledge the contributions of The STMIK Sinar Nusantara Surakarta, Indonesia, for funding this research.

References

[1] Yunita, M. *Analisis Hubungan antara Pertumbuhan Ekonomi dengan Indeks Pembangunan Manusia.* ITS. Surabaya. 2007.
[2] Klawonn, K. Fuzzy Clustering and Fuzzy Rule. *Science Journal*. 245-252. 1997.
[3] Widodo. *Perbandingan Metode Fuzzy C-Means dan Fuzzy C-Shell Clustering (Kasus Kabupaten Kota Pulau Jawa berdasarkan variabel pembentuknya)*. ITS. Surabaya. 2012.
[4] Widodo, S. Lung Field Segmentation on Computed Tomography Image using Active Shape Model. *Kursor Journal*, 7(2): 99-108. 2013.
[5] Oyelade, O. Application of K-Means Clustering Algorithm for Prediction of Students Academic Performance. *International Journal of Computer Science and Information Security*. 7(1): 292-295. 2010.
[6] Siska, S. T. Analisa dan Penerapan Data Mining untuk menentukan Kubikasi Air Terjual berdasarkan Pengelompokan Pelanggan menggunakan Algoritma K-Means Clustering. *Jurnal Teknologi Informasi dan Pendidikan*. 9: 86-93. 2016.
[7] Bunkers, J. Definition of Climate Regions in the Northen Plains using an Objective Cluster Modification Technique. *Journal of Climate*. 9: 130-146. 1996.
[8] Pethalakshmi, A. Modification in K-Means Clustering Algorithm Through Affinity Measure to Increase The Cluster Uniqueness. *International Journal on Soft Computing*. 1(2). 2012.
[9] Goryawala, M. Analyzed on 3-D Liver Segmentation using Combined Approach of K-Means and Segmentation Algorithm. *Bioinformatics and Medical Engineering*. 5(1): 7-14. 2012.
[10] Siddiqui, U. The Optimized K-Means Clustering Algorithm using An Image Segmentation with The Capability of Avoiding The Dead Centre and Trapped Centre at Local Minima. *Engineering Application of Artificial Intelligence.*13(3): 263-278. 2012.
[11] Vlase, M., Munteanu, D., and Istrate A. An Improvement of K-Means Clustering Algorithm using Various Patent Metadata. *International Journal on Computers and Mathematics*. 49: 757-763. 2012.
[12] Li, X. Evaluates Clustering in Image Indexing using K-Means Clustering Algorithm and Genetic Algorithm. *Journal of Intelligent Information Systems*. 23: 5-16. 2012.
[13] Badan Pusat Statistik (BPS). *Jawa Tengah dalam Angka*. Badan Pusat Statistik: Jawa Tengah. 2014.