Hierarchical Clustering for Development Equality Analysis: Indonesian Data of Educational Support (2011 – 2014)

Feri Wijayanto
Department of Informatics, Faculty of Industrial Technology, Universitas Islam Indonesia
feri.wijayanto@uii.ac.id

Abstract. Indonesia which contains more than 30 provinces with the decentralization system needs to identify its development equality. Because inequality in development performance will bring disparity among provinces. At present, the development monitoring is using the indicator’s values and comparing those values among provinces. There are no tools which could be used to identify the general development performance, moreover regarding the equality. This research wants to see the possibility of using hierarchical clustering to observe this equality, especially on educational support development. In result, the graph which is plotted using the dissimilarity values as a side result of hierarchical clustering could describe the trend of the equality.

1. Introduction
Indonesia is a vast country which contains more than 30 provinces which are separated by the ocean because Indonesia is an archipelago. It will need a ton efforts from the central government when all of the development is centralized. But on contrary, because of the implementation of decentralization system, the development performance need to be monitored and evaluated since every province has different problem and their development planning will not be the same to each other.

At present, the development performance is indicated by a pack of indicators. With the condition, the monitoring process is conducted by comparing the indicator’s value of every province. Those indicator’s values are published annually by the Central Bureau of Statistics or related ministry. Even though the values could be compared, the general performance of each province could not be compared [1]. Moreover, the general equality of the development of all provinces could not be identified. Start from that problem this research will try to use hierarchical clustering to evaluate the development equality in Indonesia.

The usage of clustering method is coming up along with idea that the terminology of “equality” is assumed to be the same with “similar”. It means that the performances are “equal” among all of objects if they are “similar” to each other. The clustering method which is developed by using “dissimilarity” concept which is assumed to be the same with the term of “inequality”. So by using the dissimilarity concept along with clustering method, it could be inferred that the lower the dissimilarity, it would higher in equality. Based on those backgrounds, this research plans to observe whether the dissimilarity concept in clustering method could be used to monitor the development equality, especially in educational support development.
2. Methodology

Basically, the methodology of this research, the analysis of development equality, is based on the clustering analysis. The clustering process need to be conducted before the dissimilarities could be obtained. The dissimilarity values are projected to be fetched from the clustering process because by implementing the clustering the similar objects could be grouped into a cluster and the other similar objects into another cluster. In the end, the similarity of an object relative to the other object could be concluded. Based on the research objective, the hierarchical clustering is selected to be used in this research. The usage of the hierarchical clustering due to the possibility of using dendrogram [2] to represent the creation process of cluster which is equipped with the dissimilarity values of each sequence.

In general, the methodology which are used in this research are shown in Figure 1. The first stage of this research is collecting data which are related to Indonesia educational support in four years start from 2011 to 2014. After the collection process, the clustering methodology is implemented, starting with data preprocessing which continues with normalization and transformation [3]. The stage then moved on computing the optimal number of cluster and the clustering process itself. In this research, the important results which are expected from the clustering process is not the cluster, but the distances which occurs on the cluster creation process. Then the distances are collected and represented as a graphic. Analysis and interpretation to the graphic need to be conducted in the last step in order to conclude the hierarchical clustering possibility to observe the development equality.

2.1. Educational support dataset collections

The data is collected from the Senior Secondary School Statistic which is published by The Educational Data and Statistic Center, Ministry of Education and Culture. While the dataset is formed based on the input – support – output – context model which have been proposed in previous research [4] which is shown on Figure 2. In previous research on educational effectiveness, it is proposed to use an input – process – output – context to measure the educational development effectiveness. Basically this model is adapted from Scheerens’ model [5] which using the OECD indicators. So, based on the model the educational support dataset will be formed by using the six indicators in the context component.

2.2. Clustering analysis

The clustering algorithm which will be used in this research is the simplest because the focus of the discussion here is not the resulted cluster but rather on the distance values which are produced in conjunction with clustering process. So, the hierarchical clustering method is used in this research. Another reason of the method usage is due to the requirement of distances tracking for each sequence of cluster creation. Moreover, the hierarchical clustering, especially complete link, has the ability to focus on outliers [6] which means it becomes possible to identify the unusual phenomenon.
On general, the common hierarchical clustering methodology has been used. The educational support dataset which has six features like on Table 1 is normalized and transformed before it clustered. The min-max normalization[2] and cost-benefit analysis[7] transformation is implemented on the preprocessing. The methodology for clustering analysis could be described in Figure 3. Basically, on clustering analysis, number of clusters is an important information but here, the step is excluded because the goal here is only obtaining the distance of the agglomerative process.

### Table 1. Educational Support features' dataset

| No. | Feature                              |
|-----|--------------------------------------|
| 1   | Students/School Ratio                |
| 2   | Students/Classroom Ratio             |
| 3   | Classroom/room ratio                |
| 4   | Qualified Teachers Percentage        |
| 5   | Students/Teachers Ratio              |
| 6   | Proper Classroom Percentage          |

![Figure 2. The proposed effectiveness model [4]](image)

2.3. **Distance values collections**

One other thing that makes hierarchical clustering is implemented on this research is dendrogram [8]. The cluster creation on hierarchical clustering is having the capability to be represented as a dendrogram. Dendrogram has parameter which is needed for this analysis that is known as height, as represented on Figure 4, Figure 5, Figure 6 and Figure 7. Height is parameter which represents the distance values which are computed for every agglomeration of an object. The lower the distance the more similar it should be and the more equal it becomes. Those heights are then to be collected in order to be graphed on a trendline graph in agglomeration process sequence order which is shown in
Figure 10. Further than that, the average of those heights and the highest heights in four years are taken in order to be graphed on Figure 8 and Figure 9.

3. Results and Discussions

After the distances, based on the dissimilarities from the clustering process, are collected, they are then plotted on graphs which are observable on the Figure 8 and Figure 9. Figure 8 shows the trend of dissimilarities by plotting the highest value of computed distance (last computed distance) from each year dataset while Figure 9 forms the trend by plotting the computed average of distance values from each year dataset. By comparing Figure 8 and Figure 9 it is easy to conclude that the trend is opposite. Figure 8’s trend is relatively increasing while Figure 9 is relatively declining. Basically, it could be simply inferred that on average the dissimilarities is declining but on contrary there are provinces which have quite different performance on educational support provision.

Theoretically, equality in educational development could be obtained if the dissimilarities are zero or approaching zero, but that is impossible. Actually, if it could be determined how near to zero of dissimilarities which could be said that is still considered equal, the trendline graph on Figure 8 and Figure 9 would be rich in information regarding the equality.

The most used distance formula in clustering computation is the Euclidean as written in equation (1). Based on the equation (1), if the dataset contains six parameters the highest value of distance which could be obtained is six. It is assumed that the datasets are normalized into 0 to 1. Thus, in general, what could be said on the dissimilarities by referring to Figure 8 and Figure 9 is that the
average dissimilarities is quite low that is only about 0.5 or 8% to the highest value possible for the first three years and 0.38 or 6.7% for 2014.

\[ D_{ij} = \left( \sum_{l=1}^{d} (x_{il} - x_{jl})^2 \right)^{1/2} \]  

(1)

where:
- \( D_{ij} \) = The distance between object i and j
- \( x_{il} \) = The lth parameter of object i
- \( x_{jl} \) = The lth parameter of object j

The results on the Figure 8 and Figure 9 could be clarified with Figure 10 which shows the trend of dissimilarities on the sequence of cluster creation in four years (2011-2014). For the highest value (last value) of dissimilarity in each year is 2014, 2012, 2011 and 2013 respectively from the highest value to the lowest. It matches with trend on Figure 8. If the Figure 10 is closely observed, it could be predicted although 2014 has the highest dissimilarity on the last cluster creation on contrary averagely 2014 has the lowest dissimilarity.

So basically, by analyzing the Figure 8 and Figure 9 it could be inferred that the graph could be used to describe the equality trend in a certain range of time. In this research case, it could be inferred that in general, the educational support development performance among provinces in Indonesia is becoming more equal year by year from 2011 to 2014 but on contrary the highest dissimilarity in general become higher.

![Figure 10. Dissimilarities trend of cluster sequence in four years (2011-2014)](image)

4. Conclusion

In conclusion, the plotted dissimilarities could be used to observe the growth of the dissimilarities year by year. The growth of dissimilarities, if compared to the definition, would means the decline of equality. On contrary, the equality would be obtained if only the dissimilarity is zero. Even though it would be impossible, it could be simply said that the lower the dissimilarity would be higher in equality.

Using the graph behavior, the equality of educational support development could be observed and deduced. Observing on average dissimilarity trend, it could be concluded that in general the performance of educational support development becoming more equal. While as a supplementary, by
using the highest dissimilarity value trend, it could be alerted that there are any extraordinary cases or outliers in some provinces.

5. References

[1] Wijayanto F Input-Support-Output Model Evaluation Using Clustering Analysis on Indonesia High School Dataset Adv. Sci. Lett. (forthcoming)

[2] Han J, Kamber M and Pei J 2006 Data Mining: Concepts and Techniques (San Francisco: Morgan Kaufmann Publishers Inc.)

[3] Berkhin P 2006 A Survey of Clustering Data Mining Techniques Grouping Multidimensional Data (Berlin/Heidelberg: Springer-Verlag) pp 25–71

[4] Wijayanto F Clustering Analysis on Indonesian Education Quality Performance Using Input-Output Model Adv. Sci. Lett. (forthcoming)

[5] Scheerens J, Luyten H and van Ravens J 2011 Measuring Educational Quality by Means of Indicators Perspectives on Educational Quality (Springer Netherlands) pp 35–50

[6] Madhulatha T S 2012 An Overview on Clustering Methods IOSR J. Eng. 2 719–25

[7] Larose D T 2005 Discovering Knowledge in Data: An Introduction to Data Mining (John Wiley & Sons)

[8] Kotsiantis S B and Pintelas P E 2004 Recent Advances in Clustering: A Brief Survey Methods 1 73–81