International Journal of Information Science 2017, 7(1): 12-15
DOI: 10.5923/j.ijis.20170701.02

A Decision Support System to Cluster a Priority Development Sub Town in Education Field with K-Means Clustering Algorithm (Case Study Center Java Province of Indonesia)

Nursikuwagus Agus*, Hartono Tono*

Information System, Faculty of Technic and Computer Science, Indonesia Computer University, Indonesia

Abstract Education is one field in many countries that has been supporting to help the people growth. In the knowledge manner, education is importance activity to endorse and increase the people in economic and development culture. In the sub town, the problem of government policies is choosing a priority where the sub town that has a high priority and essential to realize their development in education. The purpose of the research is applying K-Means Clustering algorithm and cluster the education data, such as population, class room, and teacher. This process has been useful to cluster the data in education field. The high priority in the system, it can be supported by government firstly. In clustering process, we have been using 35 data that has been distributed in central java. The algorithm that has processing conducted by cluster technic that includes three terms such as weak frequency (cluster 1), middle frequency (cluster 2), and tight frequency (cluster 3). So, we have been setting for K-Means value is three clusters. The conclusion of the research is the sub town that has a high priority would be endorsed in education development firstly is around Magelang with 11 districts.

Keywords K-Means, Decision Support, Clustering, Education, Development

1. Introduction

Clustering is an unsupervised technique based analysis and data mining techniques. Many studies were using this technique, to solve the problems in obtaining the desired results. (Fahmida Afrin, 2015). Past research on K-Means algorithm has been made to the customer segmentation (Fahmida Afrin, 2015). Meanwhile (Archana Singh, 2013) using the K-Means technique to divide the data into K clusters, calculation of the distance between a predetermined point become a factor in cluster point has been obtained. (Farhad Soleimanian Gharehchopogh, 2012) using the K-Means to determine infiltration activities in a computer network. (Rajagopal, 2011) was using the K-Means to cluster customers with high-profit categories on, high value and low risk to the customer. (Soumi Ghosh, 2013) using the K-Means to determine clusters of business transactions conducted by the company.

Development education is one of the main priorities in the national development agenda and was instrumental in achieving progress in many areas of life such as social, economic, political and cultural. Education is one of the strategic areas that need serious attention as a means to enhance human intelligence and skill.

The purpose of this study is to implement the K-Means algorithm to a decision that involves the total population, the number of classrooms, and the number of teachers in a region in prioritization of assistance in the field of education. As a limitation, in the boundary of the problem is the data from the local government that includes cities/regencies in Central Java. In addition, the test parameters or variables that have been used are the population of the region, the number of classrooms and teachers. The amount of data used is as many as 35 districts / cities. The expected outputs of this research are clustered several districts /cities in Central Java were considered by the government which must first be assisted in improving education.

2. Literature Review

K-Means Algorithm

K-Means or Hard C-Means clustering is basically a partitioning method applied to analyze data and treats observations of the data as objects based on locations and distance between various input data points (Soumi Ghosh,

* Corresponding author:
agus235032@yahoo.com (Nursikuwagus Agus)
tnaia74@yahoo.com (Hartono Tono)
Published online at http://journal.sapub.org/ijis

2013). Partitioning the objects into mutually exclusive clusters (K) is done by it in such a fashion that objects within each cluster remain as close possible to each other but as far as possible from objects in other clusters. (Soumi Ghosh, 2013). Each cluster is characterized by its centre point i.e. centroid. The distances used in clustering in most of the times do not actually represent the spatial istances. In general, the only solution to the problem of finding global minimum is exhaustive choice of starting points. But use of several replicates with random starting point leads to a solution i.e. a global solution (Soumi Ghosh, 2013). In a dataset, a desired number of clusters K and a set of k initial starting points, the K-Means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is the point whose coordination. The algorithm have been pursued by (Archana Singh, 2013), it can be seen at the below.

Algoritma K-Means Euclidian Distance (Archana Singh, 2013):

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1. Select 'c'cluster centers randomly.
2. Calculate the distance between each data point and cluster centers using the Euclidean distance metric as follows:

$$Dist_{XY} = \sqrt{\sum_{k=1}^{m}\left(X_{ik} - X_{jk}\right)^2} \quad (1)$$

3. Data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. New cluster center is calculated using where, 'ci' denotes the number of data points in i-th cluster.

$$V_i = \left(\frac{1}{c_i}\right)\sum_{1}^{e_i} X_i \quad (2)$$

5. The distance between each data point and new obtained cluster centers is recalculated.
6. If no data point was reassigned then stop, otherwise repeat steps from 3 to 5.

Decision Support

Decision is activity to justify the problem solving. We can be created among problem in many decisions. Some time, the decision is difficult to reach, because bulk of data and various value. In decision support system, the crucial output that has been reaching is decision which has match between problem and conclusion. In (Efraim Turban, 2005) has defined that decision support (DSS) is collaborative from personal intellectual with the computer system capability to increase the quality of decision. The output of DSS is decision, report, or rank of event that has been choosing from many data with the characteristics defined (Efraim Turban, 2005).

Clustering

Data clustering is an unsupervised data analysis and data mining technique. Hundreds of clustering algorithms have been developed by researchers. The development of clustering methods is very interdisciplinary. The contributions have been made, for example, by psychologist, biologists, statisticians, social scientists, and engineers. (Fahmida Afrin, 2015). In the clustering methods, there are many different amount of distance, such as Euclidean distance, Minkowski distance, Manhattan Distance, etc. (Archana Singh, 2013). For this research, we have used Clustering K-Means with Euclidean Distance.

3. Research Model

In (Fahmida Afrin, 2015), they have succeed implement the K-Means algorithm for clustering the customer segmentation. Another research which has succeeded too for implementation of K-Means is (Budiarti, 2006) (Navjot Kaur, 2012) (Rajagopal, 2011) (Farhad Soleimanian Gharehchopogh, 2012). In (Budiarti, 2006), she was succeeding in clustering method for prediction of student graduate. In (Navjot Kaur, 2012), he was explained about meaningful of K-Means algorithm for ranking method. (Farhad Soleimanian Gharehchopogh, 2012), clustering about intrusion in network computer system from unknown intruder. In association with research before, it is possible to implement about decision system using K-Means algorithm. We are proposed for this research model, that the population, classroom needs, and teacher needs can be parameterized model to justify the cluster, especially in development of education in Central Java. We can been seen at the figure 1, we can be shown the correlation between parameters in X,Y,Z graphics.

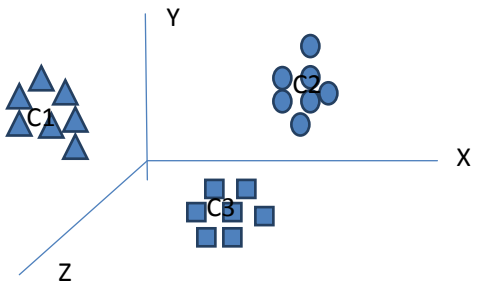

Figure 1. The Clustering Model for population (C1), classroom (C2), and teacher needs (C3)

4. Data Analysis

The process for clustering has been collected from the district government data. There are 35 districts with various parameters that consisted of population, classroom, and teacher amount. Every parameter or variable is defined as numeric value. The K-Means algorithm, have a criteria for the process such as centroid value. In that algorithm, we have been conducting the process as a sequential process for every stage. At the table 1, is shown about distribution value for every variable.

Table 1. Example data, retrieved from district government

No	District	Population	Total of Classroom	Total of Teachers
1	Banjarnegara	327472	9767	9494
2	Banyumas	410805	16795	15360
3	Batang	208766	3657	6131

Select Cluster "C" Randomly

Clustering which has defined, we have obtained the value for centroid point firstly that content of C1, C2, and C3. C1 has defined as a low priority value, C2 as a medium value, and C3, as a high priority value. We used C, it means clustering. The value of K, we have been set K = 3, because the cluster which has expected is three clusters. The algorithm that has been run with randomly was obtained that the centroid value, such as C1 = 248137.4286 means for average of population, C2 = 5950.6857 means for average of classroom, and C3 = 8637.8857 means for average of teacher. At the 35 point which has predefined, the algorithm has been selecting for the first three clusters for C1, C2, and C3.

Calculate the Distance between Each Data Point and Cluster Centers

The centroid value that has been defined is used for calculating each data with eq.1 and eq.2. The result after executed the process, we have been clustering by 35 data. The final result until no more close point into centroid value can be seen at the table 2. There are three clusters which can be mined. First cluster is the city which has distance centroid value that close with Semarang district, Second cluster is that centroid value that close with Sukoharjo district. Third cluster is that close with Magelang district. The research has been formulated that the cluster is divided into three characteristic clusters. In associate with the result, we can be concluded that the priority of the cluster is obtained near to Magelang district because has a high priority to supported by Central Java government. We can be summarized that the cluster-1 has 22 districts, cluster-2 has 2 districts, and cluster-3 has 11 districts. At the table3, it can be displayed the distribution every cluster after K-Means Algorithm processing.

5. Discussion

On the process that has been executed and model proposed, it can be stated that the model clustering can be constructed in the cases above. Population, classroom, and teacher amount can aid to determine the priority support in education field. The table 3, comparison with other cluster model. We have found that the clustering algorithm give the same result with two cluster that has predefined, especially for hierarchical clustering algorithm.

Table 2. The Centroid cluster after iteration-5th

Attribute	Full Data	C1	C2	C3
	35*	**22****	**2****	**11****
City	Magelang	Semarang	Sukoharjo	Magelang
Population	248137.4286	332078.7727	249813.5	79950
Classroom	5950.6857	7127.5455	5222	3729.4545
Teacher	8637.8857	10635.0455	968.5	6038

*Total city in Central Java
**Number of city for each cluster

Table 3. Comparison Model for Clustering Algorithm

Algorithm	Data	Population Centroid	Classroom Centroid	Teacher Centroid	Cluster1 (Totally)*	Cluster2 (Totally)*	Cluster3 (Totally)*
K-Means	35	248137.43	5950.69	8637.89	22	2	11
Filtered Clustering	35	248137.43	5950.69	8637.89	22	2	11
Hierarchical Clustering	35	248137.43	5950.69	8637.89	33	1	1
Expectation Maximization	35	151022.08	3215.2103	4389.1482	10	13	12

*The district amount

6. Conclusions

In research process above, we can be told that the K-Means algorithm has supporting by clustering the data. The parameters, which have defined, were giving the significant value for clustering, such as population, classroom, and teacher. At the end, we can be concluded that 11 districts which have a high priority to support in education field, and clustered around in Magelang. The 22 districts have clustered around Semarang in low priority and 2 districts in medium priority around in Sukoharjo.

REFERENCES

[1] Archana Singh, A. Y. (2013). K-Means with Three Different Distance Metrics. International Journal of Computer Applications, 67(10), 13-16.

[2] Budiarti, A. G. (2006). Studi Karakteristik Kelulusan Peserta Didik Dengan Teknik Clustering. Nationa Conference System and Informatics. Bali, Indonesia.

[3] Fahmida Afrin, M. A.-A. (2015). Comparative Performance of Using PCA with K-Means And Fuzzy C Means Clustering for Customer Segmentation. International Journal of Scientific & Technology Research, 4(10), 70-74.

[4] Farhad Soleimanian Gharehchopogh, N. J. (2012). Evaluation of Fuzzy K-Means and K-Means Clustering Algorithms in Intrusion Detection Systems. International Journal of Scientific & Technology Research, 1(11), 67-72.

[5] Giyanto, H. (2008). Penerapan Algoritma K-Means, K-Medoi, Gath Geva. Yogyakarta, Indonesia: Unpublished.

[6] Navjot Kaur, J. K. (2012). Efficient K-Means Clustering Algorithm Using Ranking Method in Data Mining. International Journal of Advanced Research in Computer Engineering & Technology, 85-91.

[7] Neha Aggarwal, K. A. (2012). Comparative Analysis of K-Means and Enhanced K-Means Clustering Algorithm for Data Mining. International Journal of Scientific & Engineering Research, 3(3), 1-8.

[8] Rajagopal, S. (2011). Customer Data Clustering Using Data. International Journal of Database Management Systems (IJDMS), 3(4), 1-11.

[9] Soumi Ghosh, S. K. (2013). Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. International Journal of Advanced Computer Science and Applications (IJACSA), 4(4), 35-39.

[10] Tan, P. S. (2006). Introduction to Data Mining. New York: Pearson Education.