Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province

Agus Perdana Windarto, Muhammad Noor Hasan Siregar, Wildan Suharso, Barany Fachri, Adi Supriyatna, Irmawati Carolina, Yoyon Efendi, Dafwen Toresa

STIKOM Tunas Bangsa Pematangsiantar, North Sumatera, Indonesia
Universitas Graha Nusantara, Padangsidempuan, North Sumatera, Indonesia
Universitas Muhammadiyah Malang, Indonesia
Universitas Pembangunan Panca Budi, Medan, North Sumatera, Indonesia
Universitas Bina Sarana Informatika, Indonesia
STMIK AMIK Riau, Indonesia
Universitas Lancang Kuning, Indonesia

agus.perdana@amiktunasbangsa.ac.id

Abstract. One of the important needs of environmental health is clean water. Clean water is the most important necessity of living beings in supporting survival. The study aimed to cluster the number of clean water customers by province (1995-2015). The method used is data mining clustering using k-means. The sample data used 34 provinces with attribute assessment of the number of clean water customers by province. The clustering process is done with 3 clusters, namely (C1) Cluster High, (C2) Cluster Normal and (C3) Cluster Low, for the number of clean water customers who are low on the need of clean water. The results showed, C1: 6 provinces, C2: 4 provinces and C3: 24 provinces. The end centroid values used are: C1 (296587.22), C2 (995898.56) and C3 (70832.29). The results obtained on the Davies-Bouldin index for "the number of clean water consumers" are -0.470. Based on performance results, it can be concluded that k-means algorithm is best because it has the smallest Davies-Bouldin index value. Based on research results, 70% of Indonesian people are still low awareness of the need for clean water.

1. Introduction

One of the important needs of environmental health is clean water. Clean water is the most important necessity of living beings in supporting survival. Where every day humans need clean water to drink, cook, bathe, wash and so on. In relation to clean water sources. The role of the water supply establishment in distributing clean water to households is very important considering its usefulness and its benefits. The number of water companies in the last five years tends to fluctuate. By 2015 there are 539 companies registered as clean water companies. Figure 1 shows the number of companies during 2011-2015.

Source: BPS-Statistisc Indonesia

Figure 1. Number of Water Supply Companies
The trust of the customer (household) to the water supply establishment shows the level of human awareness as increase on awareness on clean water need. Increased awareness of the need for clean water is getting higher. This is evidenced by the increasing number of clean water customers from 2011 - 2015. The average growth of the average number of customers per year is 6.42%. This is dominated by non-business customers (households and institutions) who consume clean water needs with an average volume of clean water distribution of 10.01% per year, which means non-business customers consume clean water by 64.45% of the total water distributed. This study aims to classify clean water customers by region (Province) where data is obtained from the BPS-Statistic Indonesia [1].

Data mining approach evolves to address various problems concerning data processing. Some researchers [2][3][4][5] use data mining techniques to solve the problem. Data mining has several techniques, including classification, estimation, prediction and clustering. Clustering Technique [6] is a grouping of records, observing or observing and forming classes of objects that have similarities. Cluster is a collection of records that have similarities with each other and have an unlike the records in the cluster [7][8][9]. It is expected that this research can provide input to the government in mapping the cleaned water customers by province. The result of mapping in the form of cluster becomes the end result of the research. Clustering The number of cleaned water customers subscribers is the least of concern to the government to encourage communities and water supply establishment in the province to raise awareness of clean water needs.

2. Methodology
2.1. Data Mining
Data mining is an interdisciplinary subfield of computer science. It is the use of automatic data analysis techniques to uncover previously undetected relationship among data items[10]. It allows users to analyze data from various angles and dimensions, classified it and precise the relation recognized[3], [7]–[9]. The K-means algorithm involves randomly selecting K initial centroids or mean where K is a user defined number of desired clusters. For each of the object the distance is calculated between center points and data points and with minimum distance data, cluster is generated. This data points are far from another cluster or group [8].

2.2. Clustering
Clustering is the essential aspect of data mining. It is the technique of grouping of data in different groups by of their similarity measures. It means data items in the same group that are called cluster are more similar to each other than to those in other groups. Clustering is an unsupervised learning [4]. Clustering techniques mainly used two algorithms: Hierarchical algorithm and Partition algorithm. In the hierarchical algorithm, the dataset is divided into smaller subset in a hierarchical manner whereas in partition algorithm dataset is partitioned into the desired number of sets in a single step. K-means clustering is most popular partition algorithm. It uses in many application for producing the accurate result because of its simplicity of implementation [4].

2.3. K-Means Algorithm
K means algorithm work as follows [8]:
1. Initialization by setting initial centroids with a predefined k.
2. Cluster or group the data points in given k clusters.
3. Assign data or objects to nearest cluster center as per distance function.
4. When all objects are assigned recalculate or update the position of k centroids.
5. Repeat step 3 and 4 until the centroids no longer move.

3 Result And Discussion
To perform calculations using the K-Means method on the number of cleaned water customers by province, the researcher used data from BPS-Statistic Indonesia for the number of cleaned water customers (1995-2015). The attributes used are the number of cleaned water customers (1995-2015)
with the research data used are 34 provinces. Before the data were processed using the K-Means method, the data was first normalized by taking the average value of the number of cleaned water customers (1995-2015) in each province. So obtained data normalization of cleaned water customers (1995-2015) as in the following table:

| Provinces        | Average     |
|------------------|-------------|
| Aceh             | 79211,476   |
| North Sumatra    | 573045,43   |
| West Sumatra     | 170456,52   |
| Riau             | 81655,238   |
| Jambi            | 124316,33   |
| South Sumatra    | 220406,95   |
| Bengkulu         | 45126,19    |
| Lampung          | 69946,381   |
| Bangka Islands   | 13042,667   |
| Riau Islands     | 101336,19   |
| DKI Jakarta      | 1029416,7   |
| West Java        | 900283,9    |
| Central Java     | 918341,29   |
| DI Yogyakarta    | 104699,62   |
| East Java        | 113555,24   |
| Banten           | 133464,48   |
| Bali             | 287000,9    |

Based on table 1, the data clustering is done using one assessment criterion, namely the average number of cleaned water customers (1995-2015). The clustering is done on 3 clusters, namely: (C1) Cluster High, for the number of cleaned water customers who are high on the need of clean water, (C2) Cluster Normal, for the number of cleaned water customers who are normal on the need of clean water and (C3) Cluster Low, for the number of cleaned water customers who are low on the need of clean water. At this stage, data analysis of the average number of clean water customers in Indonesia will be processed using tools RapidMiner.

### 3.1. Centroid Data

In the application of K-means algorithm using tools rapidminer, the value of centroid comes from the grouping of data performed. In this case, the researchers conducted 3 clusters. The cluster point determination is done by taking the highest value for Cluster High (C1), the average value for Normal Cluster (C2) and the smallest value for Cluster Low (C3). The initial centroid data for each cluster can be seen in the table below:

| Data Cluster        | Number of cleaned water customers |
|---------------------|-----------------------------------|
| (C1) Cluster High   | 1135552                           |
| (C2) Cluster Normal | 219502,7                          |
| (C3) Cluster Low    | 5016,143                          |

### 3.2. K-Means using tools RapidMiner

In clustering the number of cleaned water customers by province, the authors use K-means algorithm to cluster data based on attributes at cluster center distance as in table 2. Iteration is the process of execution on K-Means to classify data based on cluster center of distance value. The cluster center distance value will continue to change according to the number of iterations performed. The K-Means process will continue to zero in until the data clustering is the same as the previous iteration data cluster. In other words, the process will continue iterating until the data in the last iteration is the same as the previous iteration. Here is the architecture design with Rapidminer tools.
Figure 2. Architecture of K-Means

Figure 2 describes a model first trained and read through Read Excel using excel data; information related to Read Excel is studied by the model. Then the model can be applied to other Read Excel normally for prediction. All required parameters are stored in the model object. It is mandatory that both Read Excel should have the exact same number, order, type, and role attribute. If the meta data property of Read Excel is inconsistent, this can cause serious errors. Based on the design of Figure 2, the rapidminer tools will cluster the 'the number of cleaned water customers' area based on the input data provided. The results of the final clustering can be seen in the picture below:

Cluster Model

Cluster 0: 24 items
Cluster 1: 4 items
Cluster 2: 6 items
Total number of items: 34

Figure 3. Clustering of K-Means

Figure 3 describes the clustering of ‘the number of cleaned water customers’ by province that 6 provinces in (C1) Cluster High, 4 provinces in (C2) Cluster Normal and 24 provinces in (C3) Cluster Low. Based on research results, almost 70% of the number of cleaned water customers are in cluster C3. Here is a clustering chart for the number of cleaned water customers.

3.3. K-Means Performance Accuracy

To see k-means accuracy using tools RapidMiner, one of the parameters used in RapidMiner tools is used. One of them is %Performance. %Performance is used for performance evaluation of centroid-based grouping methods. This operator provides a list of performance criteria values based on the cluster centroid. The %Performance measurement parameters are avg._within_centroid_distance and davies_bouldin. The results of this study obtained the accuracy of data that excellent with the value: a) avg._within_centroid_distance = -5291292289.031 and b) davies_bouldin = -0.470. The results of k-means accuracy using tools RapidMiner can be seen in the picture below.
4. Conclusion

To cluster data on the number of cleaned water customers by k-means method using tools RapidMiner, the result of clustering with 3 clusters: C1: 6 provinces: North Sumatera, South Sumatera, Bali, South Kalimantan, East Kalimantan and South Sulawesi; C2: 4 provinces : DKI Jakarta, West Java, Central Java and East Java; C3: 24 provinces: Aceh, West Sumatra, Riau, Jambi, Bengkulu, Lampung, Bangka Island, Riau Island, Yogyakarta, Banten, West Nusa Tenggara, Central Nusa Tenggara, West Kalimantan, Central Kalimantan, North Sulawesi, North Sulawesi, Central Sulawesi, Southwest Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua and Papua.

References

[1] Kepala Badan Pusat Statistik Republik Indonesia, Water Supply Statistics. BPS-Statistics Indonesia.
[2] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, “Application of k Means Clustering algorithm for prediction of Students Academic Performance,” Int. J. Comput. Sci. Inf. Secur., vol. 7, no. 1, pp. 292–295, 2010.
[3] S. Kumar and S. K. Rathi, “Performance Evaluation of K-Means Algorithm and Enhanced Mid-point based K-Means Algorithm on Mining Frequent Patterns,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 4, no. 10, pp. 545–548, 2014.
[4] A. Yadav and S. Dhingra, “An Enhanced K-Means Clustering Algorithm to Remove Empty Clusters,” IJEDR, vol. 4, no. 4, pp. 901–907, 2016.
[5] N. Aggarwal, K. Aggarwal, and K. Gupta, “Comparative Analysis of K-means and Enhanced K-means Clustering Algorithm for Data Mining,” Int. Journai Sci. Eng. Res., vol. 3, no. 3, 2012.
[6] T. M. Kodinariya and P. R. Makwana, “Review on determining number of Cluster in K-Means Clustering,” International Journal of Advance Research in Computer Science and Management Studies, 2013.
[7] U. R. Raval and C. Jani, “Implementing and Improvisation of K-means Clustering,” Int. J. Comput. Sci. Mob. Comput., vol. 5, no. 5, pp. 72–76, 2016.
[8] M. K. Arzoo, A. Prof, and K. Rathod, “K-Means algorithm with different distance metrics in spatial data mining with uses of NetBeans IDE 8 . 2,” Int. Res. J. Eng. Technol., vol. 4, no. 4, pp. 2363–2368, 2017.
[9] A. P. Windarto, “Implementation of Data Mining on Rice Imports by Major Country of Origin Using Algorithm Using K-Means Clustering Method,” Int. J. Artif. Intell. Res., vol. 1, no. 2, pp. 26–33, 2017.
[10] B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad, “Classification of natural disaster prone areas in Indonesia using K-means,” Int. J. Grid Distrib. Comput., vol. 11, no. 8, pp. 87–98, 2018.