Implementation of Data Mining in Grouping Percentage of Blind Letters Age 15+ By Province Using K-Means Algorithm

Saifullah\textsuperscript{1*}, Nani Hidayati\textsuperscript{2}, Solikhun\textsuperscript{3}

\textsuperscript{1,2}STIKOM Tunas Bangsa, Information Systems Study Program, Pematangsiantar, North Sumatra, Indonesia.
\textsuperscript{3}AMIK Tunas Bangsa Pematangsiantar, Informatics Management Study Program, North Sumatra, Indonesia
E-mail: saifullah@amiktunasbangsa.ac.id

Abstract. Data Mining is a method that is often needed in large-scale data processing, so data mining has important access to the fields of life including industry, finance, weather, science and technology. In data mining techniques there are methods that can be used, namely classification, clustering, regression, variable selection, and market basket analysis. Illiteracy is one of the factors that hinder the quality of human resources. One of the fundamental things that must be fulfilled to improve the quality of human resources is the eradication of illiteracy among the community. The purpose of this research is to determine the clustering of illiterate communities. The results of the study are illiterate data clustering according to the age proportion of 15+, namely 1 node of the high group, the low group has 23 nodes, and the medium group of 10 nodes. The results of this study become input for the government to determine illiteracy eradication policies in Indonesia based on provinces.

1. Introduction

Data mining applications frequently involve categorical data. The traditional approach to converting categorical data into numeric values does not necessarily produce meaningful results in the case where categorical domains are not ordered. The k-modes algorithm in this paper removes this limitation and extends the k-means paradigm to categorical domains whilst preserving the efficiency of the k-means algorithm\cite{1}.

The k-means algorithm can be thought of as a gradient descent procedure, which begins at starting cluster centroids, and iteratively updates these centroids to decrease the objective function. The k-means always converge to a local minimum. The particular local minimum found depends on the starting cluster centroids. The problem of finding the global minimum is NP-complete. The k-means algorithm updates cluster centroids till local minimum is found\cite{2}.

Illiteracy is one of the factors that hinder the quality of human resources. One of the fundamental things that must be met to improve the quality of human resources is the eradication of illiteracy among the people. The purpose of this study is determines clustering of illiterate communities. The results of this study become input to the government to determine the policy of eradicating illiteracy in Indonesia based on provinces.
2. Research Methodology
Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data. Data mining has been applied in a great number of fields, including retail sales, bioinformatics, and counter-terrorism. Data mining can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Alternatively, it can be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns. Data mining is not new—it has been used intensively and extensively by financial institutions, for credit scoring and fraud detection; marketers, for direct marketing and cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling[3].

Clustering, the unsupervised classification of patterns into groups, is one of the most important tasks in exploratory data analysis. Primary goals of clustering include gaining insight into data (detecting anomalies, identifying salient features, etc.), classifying data, and compressing data. Clustering has a long and rich history in a variety of scientific disciplines including anthropology, biology, medicine, psychology, statistics, mathematics, engineering, and computer science. As a result, numerous clustering algorithms have been proposed since the early 1950s[4].

Clustering algorithms are generally used in an unsupervised fashion. They are presented with a set of data instances that must be grouped according to some notion of similarity. The algorithm has access only to the set of features describing each object; it is not given any information (e.g., labels) as to where each of the instances should be placed within the partition[5].

The k-means algorithm [6, 7, 8, 9, 10] is effective in producing clusters for many practical applications. But the computational complexity of the original k-means algorithm is very high, especially for large data sets. Moreover, this algorithm results in different types of clusters depending on the random choice of initial centroids. Several attempts were made by researchers for improving the performance of the k-means clustering algorithm.

This section describes the original k-means clustering algorithm. The idea is to classify a given set of data into k number of disjoint clusters, where the value of k is fixed in advance. The algorithm consists of two separate phases: the first phase is to define k centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Euclidean distance is generally considered to determine the distance between data points and the centroids[11].

3. Results And Discussion
3.1. Data Analysis
Data analysis is a method used to answer research problems through management procedures, especially problems related to research. The data obtained for this study were 15+ illiterate presentations from 2011-2018.

1. Determine a Centroid value that is determined manually or randomly drawn from the data.

| Table 1. Centroid Value. |
|--------------------------|
| Centroid     | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |
| Centroid 1    | 35.47| 34.7 | 32.69| 29.22| 29.2 | 28.98| 26.11| 23.21|
| Centroid 2    | 11.31| 10.08| 9.16 | 7.44 | 7.23 | 7.18 | 7.1  | 7.02 |
| Centroid 3    | 0    | 0    | 0    | 0    | 5.01 | 4.95 | 4.86 | 4.82 |
2. Calculate the distance from centroid
To calculate the distance between the centroid point and the point of each object using Euclidion Distance

\[ D_{(i,j)} = \sqrt{(X_{i1} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + ... + (X_{ki} - X_{kj})^2} \]

Then the calculation for the distance from the 1st Centroid is as follows:

\[ D_{x1,c1} = \sqrt[(4,37-35,47)^2 + (3,96-34,7)^2 + (3,34-32,69)^2 + (2,58-29,22)^2 + (2,37-29,2)^2} + (2,26-28,98)^2 + (2,06-26,11)^2 + (1,97-23,21)^2 = 77,11 \]
\[ D_{x2,c1} = \sqrt[(3,22-35,47)^2 + (2,69-34,7)^2 + (2,19-32,69)^2 + (1,43-29,22)^2 + (1,32-29,2)^2} + (1,12-28,98)^2 + (1,11-26,11)^2 + (0,93-23,21)^2 = 80,26 \]

And so on up to \( D_{x34,c1} \). Furthermore the calculation for the distance from the 2nd Centroid is as follows: And so on up to \( D_{x34,c1} \). Furthermore the calculation for the distance from the 2nd Centroid is as follows:

\[ D_{x1,c2} = \sqrt[(4,37-11,31)^2 + (3,96-10,08)^2 + (3,34-7,44)^2 + (2,58-7,23)^2 + (2,37-7,18)^2} + (2,26-7,1)^2 + (2,06-26,11)^2 + (1,97-7,02)^2 = 15,55 \]
\[ D_{x2,c2} = \sqrt[(3,22-11,31)^2 + (2,69-10,08)^2 + (2,19-9,16)^2 + (1,43-7,44)^2 + (1,32-7,23)^2} + (1,12-7,18)^2 + (1,11-7,1)^2 + (0,93-7,02)^2 = 18,69 \]

And so on up to \( D_{x34,c2} \). Furthermore the calculation for the distance from the 3rd Centroid is as follows:

\[ D_{x1,c3} = \sqrt[(4,3-0)^2 + (3,96-0)^2 + (3,34-0)^2 + (2,58-0)^2 + (2,37-5,01)^2 + (2,26-4,95)^2} + (2,06-4,86)^2 + (1,97-4,82)^2 = 9,10 \]
\[ D_{x2,c3} = \sqrt[(3,22-0)^2 + (2,69-0)^2 + (2,19-0)^2 + (1,43-0)^2 + (1,32-5,01)^2 + (1,12-4,95)^2} + (1,11-4,86)^2 + (0,93-4,82)^2 = 9,05 \]

And so on up to \( D_{x34,c3} \). So we get the distance table from Centroid and find the minimum value of the three Centroids. The Distance Tables from Centroid are as follows:
Table 2. Distance of the 1st Iteration Centroid.

|    | C1   | C2   | C3   | Centroid Distance |
|----|------|------|------|-------------------|
| 1  | 77.11| 15.55| 9.10 | 9.10              |
| 2  | 80.26| 18.69| 9.05 | 9.05              |
| 3  | 79.44| 17.87| 9.57 | 9.57              |
| 4  | 80.99| 19.44| 9.05 | 9.05              |
| 5  | 77.30| 15.72| 9.45 | 9.45              |
| 6  | 79.06| 17.50| 8.90 | 8.90              |
| 7  | 76.60| 15.02| 9.61 | 9.61              |
| 8  | 74.40| 12.84| 9.73 | 9.73              |
| 9  | 76.73| 15.15| 9.23 | 9.23              |
| 10 | 80.63| 19.06| 8.66 | 8.66              |
| 11 | 83.60| 22.06| 9.31 | 9.31              |
| 12 | 77.79| 16.22| 9.50 | 9.50              |
| 13 | 62.94| 1.52 | 18.56| 1.52              |
| 14 | 67.08| 5.47 | 15.09| 5.47              |
| 15 | 58.64| 3.09 | 22.19| 3.09              |
| 16 | 76.84| 15.30| 8.59 | 8.59              |
| 17 | 61.64| 0.00 | 19.71| 0.00              |
| 18 | 44.82| 16.91| 35.36| 16.91             |
| 19 | 58.60| 3.20 | 22.19| 3.20              |
| 20 | 61.90| 1.61 | 18.86| 1.61              |
| 21 | 80.59| 19.01| 9.26 | 9.26              |
| 22 | 78.16| 16.57| 9.43 | 9.43              |
| 23 | 80.10| 18.53| 9.11 | 9.11              |
| 24 | 79.61| 19.71| 0.00 | 0.00              |
| 25 | 83.67| 22.12| 9.55 | 9.55              |
| 26 | 75.32| 13.76| 10.59| 10.59             |
| 27 | 58.27| 3.47 | 22.53| 3.47              |
| 28 | 65.97| 4.39 | 15.98| 4.39              |
| 29 | 77.37| 15.83| 10.59| 10.59             |
| 30 | 60.62| 1.41 | 20.95| 1.41              |
| 31 | 80.37| 18.80| 9.41 | 9.41              |
| 32 | 78.96| 17.37| 9.56 | 9.56              |
| 33 | 73.39| 11.88| 11.93| 11.88             |
| 34 | 0.03 | 61.63| 79.60| 0.03              |
| 35 | 77.11| 15.55| 9.10 | 9.10              |
| 36 | 80.26| 18.69| 9.05 | 9.05              |

3. Determine Clusters or Groupings

In determining the Cluster by finding the Cluster value based on the minimum value of the Cluster value and placed in the Cluster that corresponds to the minimum value in Iteration 1. Following the Cluster table in Iteration 1 as follows:

Then proceed to the next iteration until the cluster value is the same or does not change in the next iteration cluster, the calculation is stopped.
Table 3. Distance of the 1st Iteration Centroid.

|   | C1 | C2 | C3 |
|---|----|----|----|
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |
|   | 1  | 1  | 1  |

3.2. Display of RapidMiner

To get the results of grouping in the next step, do it by clicking on the blue arrow in the top center position on the toolbar. This stage will display the final results as well as the final step in using rapidminer tools. Can be seen in Figure 1:

Information:

1. Number of Clusters 0 (Height) amounting to 1 Items
2. Number of Clusters 1 (Low) totaling 23 Items
3. Number of Clusters 2 (Medium) totaling 10 Items
4. The total number of items is 34
Figure 1. Cluster Model Values

So we can know the results of grouping rapidminer can be seen in the picture as follows:

4. Conclusion
The conclusions that can be drawn on data mining techniques in clustering Percentage of Illiteracy Age 15+ by Province with this k-means algorithm are as follows:

1. The k-means clustering algorithm can help researchers group the Percentage of Illiteracy Age 15+ by Province in Indonesia. From the percentage of illiteracy age 15+ in Indonesia, it can be seen that there are 1 province with high level clusters, namely: Papua, 10 provinces with medium level clusters namely Awa Tenga, DI Yogyakarta, East Java, Bali, NTB, NTT, West Kalimantan, South Sulawesi, Central Sulawesi, West Sulawesi, and 23 other provinces with low clusters.

2. This rapidminer application can help researchers group the Percentage of Illiteracy Age 15+ by Province in Indonesia. The manual calculation of the k-means clustering algorithm in grouping the Illiteracy Percentage Age 15+ and the application in Rapidminer show the same results.

References
[1] Oheneba-Sakyi Y, 1992 Determinants Of Current Contraceptive Use Among Ghanaian Women At The Highest Risk Of Pregnancy J. Biosoc. Sci. 24, 4 p. 463–475.
[2] Oyelade O J Oladipupo O O and Obagbuwa I C, 2010 Application of k Means Clustering algorithm for prediction of Students Academic Performance 7 p. 292–295.
[3] Koh H C and Tan G, 2005 Data mining applications in healthcare. J. Healthc. Inf. Manag. 19, 2 p. 64–72.
[4] Celebi M E Kingravi H A and Vela P A, 2013 A comparative study of efficient initialization methods for the k-means clustering algorithm Expert Syst. Appl. 40, 1 p. 200–210.
[5] Wagstaff K Cardie C Rogers S and Schrödl S, 2001 Constrained K-means Clustering with Background Knowledge Int. Conf. Mach. Learn. ICML pages p. 577–584.
[6] Nazeer K A A and Sebastian M P, 2009 Improving the Accuracy and Efficiency of the k-means Clustering Algorithm Proc. World Congr. Eng. I, May p. 6.
[7] Margaret H. Dunham, 2006 Data Mining- Introductory and Advanced Concepts, Pearson Education.
[8] McQueen J, 1976 Some methods for classification and analysis of multivariate observations, Proc. 5th Berkeley Symp. Math. Statist. Prob. (1):281–297.
[9] Pang-Ning Tan, 2007 Michael Steinback and Vipin Kumar, Introduction to Data Mining, Pearson Education.
[10] Stuart P. Lloyd, Least squares quantization in pcm, IEEE Transactions on Information Theory. 28(2): 129-136.

[11] K. A. A. Nazeer and M. P. Sebastian, 2009 Improving the Accuracy and Efficiency of the k-means Clustering Algorithm Proc. World Congr. Eng. vol. I, no. May, p. 6.