TELKOMNIKA, Vol.16, No.3, June 2018, pp. 1351~1357
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v16i3.8910

K Means Clustering and Meanshift Analysis for Grouping the Data of Coal Term in Puslitbang tekMIRA

Rolly Maulana Awangga*[1], Syafrial Fachri Pane[2], Khaera Tunnisa[3], Iping Supriana Suwardi[4]
[1,2,3]Program in Informatics Engineering, Politeknik Pos Indonesia, Indonesia
[4]School of Electrical Engineering and Informatics, Bandung Institute of Technology, Indonesia
*Corresponding author, e-mail: awangga@poltekpos.ac.id

Abstract

Indonesian government agencies under the Ministry of Energy and Mineral Resources have problems in classifying data dictionary of coal. This research conduct grouping coal dictionary using K-Means and MeanShift algorithm. K-means algorithm is used to get cluster value on character and word criteria. The last iteration of Euclidian distance calculation data on k-means combine with Meanshift algorithm. The meanshift calculates centroid by selecting different bandwidths. The result of grouping using k-means and meanshift algorithm shows different centroid to find optimum bandwidth value. The data dictionary of this research has sorted in alphabetically.

Keywords*: coal dictionary, clustering, K-means, meanshift*

1. Introduction

Data Mining is a process of extracting data or filtering data by utilising a collection of a big data. The data through a series of operations to obtain valuable information from the data. According to Daryl Pregibon, the composition of artificial statistics, artificial intelligence, and database research called data mining [1]. Data mining techniques used to explore the data using classification, prediction, grouping, outlier detection, association rules, sequence analysis, time series analysis and text mining, and some new techniques such as social networking analysis and sentiment analysis [2]. The center of research and development mineral and coal technology (Puslitbang tekMIRA) is an Indonesia government institution under the Ministry of Energy and Mineral Resources. Puslitbang tekMIRA has a dictionary of coal. This coal dictionary has one thousand coal terms. It takes time between three to five minutes to find a term inside a coal dictionary.

Based on these problems, data mining method is used to classify data dictionary. This research conduct grouping data dictionary of a coal term using data mining method. K-Means and Meanshift Algorithm were chosen in this research. K-Means algorithm was used to categories students with skills such as cognitive, communication and relational[3], to evaluate student achievement levels for course content [4], and to group data based on user information created on SNS and recommend to users in the future [5]. The grouping data based on user sentences by utilising regularity among data pursued by the user using K-Means Algorithm [6]. The genetic algorithm and K-Means are used to calculate clustered centroid with heterogeneous populations that lead to better results than using random numbers [7]. The Meanshift algorithm is used to accurately detect the location of target tracking [8] by using this method and then make it easy the accuracy of the calculation of the tracking results [9]. The Meanshift algorithm is also used to solve facial-detection and tracking-based systems [10], if an imbalance occurs, it will affect the performance of [11].

In this research the algorithm grouping a coal term in a dictionary based on character and word in a cluster. The cluster serve data dictionary of coal based on predetermined criteria. The result data of clustering using K-Means and Meanshift algorithm is shown using the matloplib plot. Matplotlib is a Python package for Plotting that produces quality production graphs. Matplotlib is designed to be able to create simple and complex plots with multiple commands [12].

Received September 18, 2017; Revised February 8, 2018; Accepted April 5, 2018

2. Research Method

2.1. K-Means

K-Means Clustering is a grouping of data, where the data in K Means Clustering K is the amount of data or the number of constants. Means is the average value of the data set as Cluster [7]. So K-Means is a method to analyse data or called data mining method where this data modelling method without using supervise or unsupervised method and K-Means is a method to classify data by using partition system. The K-Means method solves large amounts of data in groups, and the data has the same characteristics as other data. Also, the team feature also has a feature [13]. Similitude is some matrix used for similarities between instances in a cluster. K-Means is an algorithm used to generate k clusters from a collection of data sets in a simple way [14]. Algorithm 1 is an explanation of k-means implementation. Algorithm 1: k-means clustering: Randomly select k cluster centers c1, ... ck, Repeat, Set each data entity to the closest cluster center ci, Change the cluster center with the average cluster i, until the cluster center does not change[15]. K - Means algorithm formula d = distance, j = amount of data, c = centroid, x = data, c = centroid 1 . The Euclidean distance formula is described in Equation 2.

$$d(X_j, C_j) = \sqrt{\sum_{i=1}^{n} (X_j - C_j)^2} \quad (1)$$

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2} \quad (2)$$

D (*i, j*) is the distance of data between i and center cluster j. $X_k i$ is the data to i on attribute data to k. X_{kj} is the center point to j at attribute to k.

Recalculate cluster center with current cluster membership. The cluster center is the average of all data/objects in a particular cluster. If desired it can also use the median of the cluster. So mean (mean) is not the only size that can use.Reassign each object using the new cluster center. If the cluster center does not change again, then the clustering process is complete. Alternatively, return to step number 3 until the center of the cluster does not change anymore [16].

2.2. Meanshift

The average shift method is a method for determining the maximum function limit of density with separate data from a function. The average shift use as a media grouping method, in which each mode represents each group [17]-18]. The average shift method classifies data in its search mode that directs and moves data to the point region along with the iteration of the data environment built with the Gaussian Kernel [19]. Starting from one data point and iteratively will improve the approximate mode [20]. The Gaussian kernel is a differentiated multivariate kernel function used for actual calculations in assumptions [21]. Bandwidth is a free parameter that shows the effect on the estimated density generated.

3. Results and Analysis

Grouping of data of coal dictionary using K-Means and Meanshift Algorithm. The grouping results are displayed using matplotlib plot. The sample data taken are vowel data (A, I, U, E, O) for the coal dictionary. Calculation data using K-Means algorithm. The result data of the last iteration, i.e. data that has been stable and without changes.

3.1. Result K-Means

3.1.1. The letter A

The data of the last iteration calculation using euclidian distance k- means for letter A. Clustering data using character and word criteria and use the centroid value of the highest and lowest values on the criteria. In table 1 is a cluster value that has a fixed and unchanged cluster.

Table 1. Data cluster of the letter A

Character	Word
19.04536	3.365126
12.0041	3.78689
7.033561	8.84361
17.00289	1.21311
.....	.....
8.006149	7.78689
11.00447	4.78689
27.20259	11.60749

3.1.2. The letter I

The data of the last iteration calculation using euclidian distance k- means for letter I. Clustering data using character and word criteria and use the centroid value of the highest and lowest values on the criteria. In Table 2 is a cluster value that has a fixed and unchanged cluster.

Table 2. Data cluster of the letter I

Character	Word
56.4358	1.234823
62.39391	4.774678
37.21559	20.42897
43.18565	14.5126
.....	.....
59.4138	1.777967
67.47592	9.838576

3.1.3. The letter U

The data of the last iteration calculation using euclidian distance k- means for letter U. Clustering data using character and word criteria and use the centroid value of the highest and lowest values on the criteria. In Table 3 is a cluster value that has a fixed and unchanged cluster.

Table 3. Data cluster of the letter U

Character	Word
0	6.297319
24.02082	17.96443
24.02082	17.96443
.....	.....
29.01724	22.93403
23.02173	16.97265

3.1.4. The letter E

The data of the last iteration calculation using euclidian distance k- means for letter E. Clustering data using character and word criteria and use the centroid value of the highest and

lowest values on the criteria. In Table 4 is a cluster value that has a fixed and unchanged cluster.

Table 4. Data Cluster of the Letter E

Character	Word
13.15295	1.009718
14.14214	1.973457
14.14214	1.973457
.....	.....
7.071068	5.100934
10	2.625744

3.1.5. The letter O

The data of the last iteration calculation using euclidian distance k- means for letter O. Clustering data using character and word criteria and use the centroid value of the highest and lowest values on the criteria. In Table 5 is a cluster value that has a fixed and unchanged cluster.

Table 5. Data Cluster of the Letter O

Character	Word
12.33333	6.342096
3.33333	2.687423
.....	.....
2.33333	3.726783
14.66667	20.73377

3.2. Meanshift

3.2.1. The letter A

In Figure 1 is the plot result for the data of the letter A. The data used is the term data of coal letter A. The value used by the data of the term A to get bandwidth one on the plot meanshift is at point 6.76.

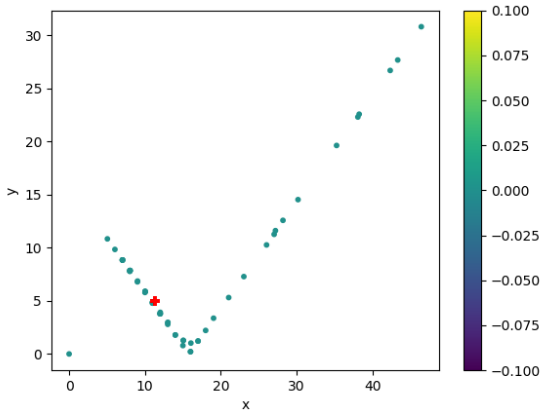

Figure 1. Plot Mean Shift letter A

3.2.2. The letter I

In Figure 2 is the plot result for the data of the letter I. The data used is the term data of coal letter I. The value used by the data of the term I to get bandwidth one on the plot meanshift is at point 18.95.

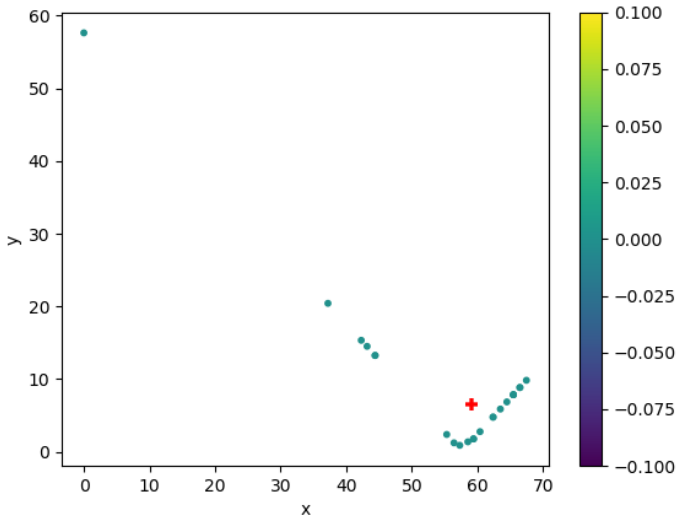

Figure 2. Plot MeanShift letter I

3.2.3. The letter U

In Figure 3 is the plot result for the data of the letter U. The data used is the term data of coal letter U. The value used by the data of the term I to get bandwidth one on the plot meanshift is at point 5.75.

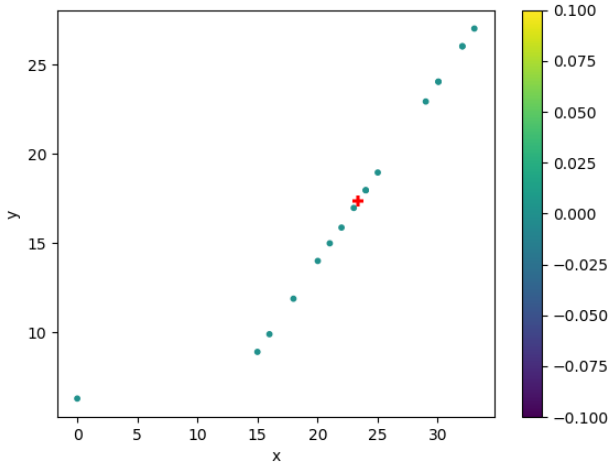

Figure 3. Plot MeanShift letter U

3.2.4. The letter E

In Figure 4 is the plot result for the data of the letter E. The data used is the term data of coal letter E. The value used by the data of the term I to get bandwidth one on the plot meanshift is at point 3.36.

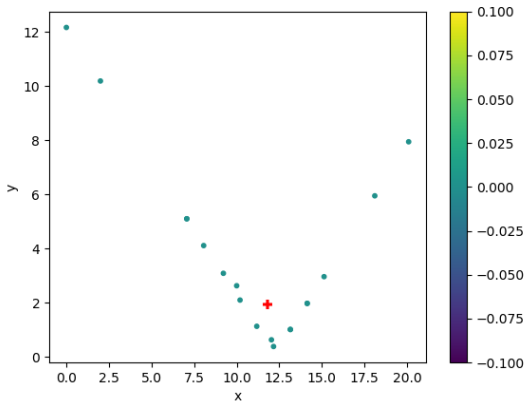

Figure 4. Plot MeanShift letter E

3.2.5. The letter O

In Figure 5 is the plot result for the data of the letter O. The data used is the term data of coal letter O. The value used by the data of the term I to get bandwidth one on the plot meanshift is at point 7.

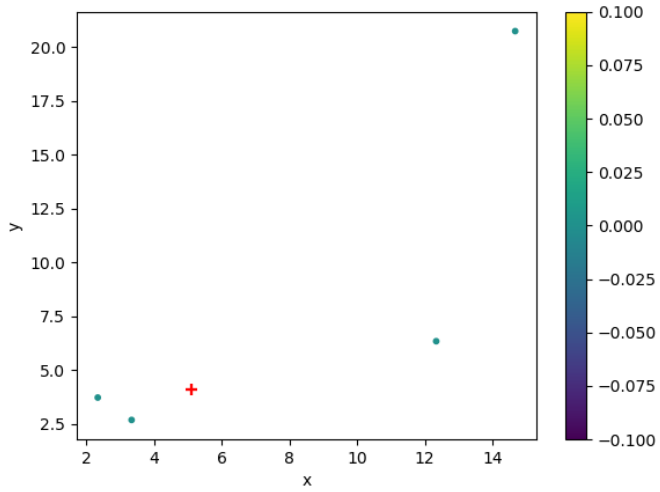

Figure 5. Plot MeanShift letter O

4. Conclusion

Grouping a coal dictionary using the k-means algorithm produces a cluster value on character and word criteria. The last iteration of the Euclidian distance calculation results in different cluster values in each alphabet. The centroid value of the k-means calculation is then combined with the MeanShift algorithm. Centroid calculation results of the meanshift algorithm result in different bandwidth. Letter A produces a bandwidth of 6.76, for the letter I produce bandwidth 18.95, letter U produces bandwidth 5.75, and letter E produces bandwidth 3.36. The grouping of coal dictionaries in this study facilitates for a group of terms in the letter. Suggestions for future research, can classify the data dictionary of coal for the second, third and subsequent letters, and combine other algorithms in grouping such as K-Nearest Neighbour (KNN) algorithm.

References

[1] F Gorunescu, Data Mining: Concepts, models and techniques. *Springer Science & Business Media*, 2011; 12.

[2] MI Ramadhan et al. *An analysis of natural disaster data by using k-means and k-medoids algorithm of data mining techniques.* in Quality in Research (QiR): International Symposium on Electrical and Computer Engineering, 2017 15th International Conference on. IEEE. 2017: 221–225.

[3] J Yi, S Li, M Wu, HA Yeung, WW Fok, Y Wang, F Liu. *Cloud-based educational big data application of apriori algorithm and k-means clustering algorithm based on students' information. in Big Data and Cloud Computing (BdCloud),* 2014 IEEE Fourth International Conference on. IEEE. 2014: 151–158.

[4] L Guoli, W Tingting, Y Limei, L Yanping, G Jinqiao. *The improved research on k-means clustering algorithm in initial values.* in Mechatronic Sciences, Electric Engineering and Computer (MEC), Proceedings 2013 International Conference on. IEEE. 2013; 2124–2127.

[5] SH Jung, JC Kim, CB Sim. Prediction data processing scheme using an artificial neural network and data clustering for big data. *International Journal of Electrical and Computer Engineering (IJECE)*. 2016; 6(1): 330-336.

[6] A Bedboudi, C Bouras, MT Kimour. An heterogeneous population-based genetic algorithm for data clustering. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*. 2017; 5(3): 275–284.

[7] Q Shi, L Xu, Z Shi, Y Chen, Y Shao. *Analysis and research of the campus network user's behavior based on k-means clustering algorithm.* in Digital Manufacturing and Automation (ICDMA), 2013 Fourth International Conference on. IEEE, 2013: 196–201.

[8] Y Cheng, W Hongyu, W Xiaohong. *Positioning method research for unmanned aerial vehicles based on meanshift tracking algorithm.* in Control and Decision Conference (CCDC), 2017 29th Chinese. IEEE 2017: 989–994.

[9] RM Awangga, NS Fathonah, TI Hasanudin. *Colenak: GPS tracking model for post-stroke rehabilitation program using AES-CBC URL encryption and QR-Code.* in 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE). IEEE. Nov 2017: 255–260. [Online]. Available: http://ieeexplore.ieee.org/document/8285506/

[10] A Salhi, Y Moresly, F Ghozzi, A Fakhfakh. *Face detection and tracking system with block-matching, meanshift and camshift algorithms and kalman filter.* in Sciences and Techniques of Automatic Control and Computer Engineering (STA), 2017 18th International Conference on. IEEE. 2017: 139–145.

[11] IN Yulita, S Purwani, R Rosadi, RM Awangga. *A quantization of deep belief networks for long short-term memory in sleep stage detection.* in Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), 2017 International Conference on. IEEE, 2017: 1-5.

[12] N Ari, M Ustazhanov. *Matplotlib in python.* in 2014 11th International Conference on Electronics, Computer and Computation (ICECCO). IEEE, sep 2014. [Online]. Available: https://doi.org/10.1109/icecco.2014.6997585

[13] J MacQueen et al. *Some methods for classification and analysis of multivariate observations* in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Oakland, CA, USA. 1967; 1(14): 281–297.

[14] AK Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters.* 2010; 31(8): 651-666.

[15] Z Gheid, Y Challal. Efficient and privacy-preserving k-means clustering for big data mining. *in Trustcom/BigDataSE/I SPA, IEEE*. 2016: 791–798.

[16] K Obbie. Penerapan algoritma klasifikasi data mining id3 untuk menentukan penjurusan siswa sman6 semarang. Skripsi, Fakultas Ilmu Komputer, 2014.

[17] Y Liu, SZ Li, W Wu, R Huang. Dynamics of a mean-shift-like algorithm and its applications on clustering. *Information Processing Letters.* 2013; 113(1): 8–16.

[18] G Chen, Q Chen, D Zhang. *Mean shift: A method for measurement matrix of compressive sensing*. in Digital Home (ICDH), 2014 5th International Conference on. IEEE. 2014: 64–69.

[19] YA Ghassabeh. A sufficient condition for the convergence of the mean shift algorithm with gaussian kernel. *Journal of Multivariate Analysis.* 2015; 135: 1–10.

[20] K Fukunaga, L Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory.* 1975; 21(1): 32–40.

[21] C Yang, R Duraiswami, NA Gumerov, L Davis. Improved fast gauss transform and efficient kernel density estimation. *in null. IEEE.* 2003: 464.