A LOF K-Means Clustering on Hotspot Data

R R Muhima[1*], M Kurniawan[2], O T Pambudi [3]
[1,2,3]Teknik Informatika ITATS, Arief Rachman Hakim 100, Surabaya, Indonesia
[1]ranimuhima@itats.ac.id*; [2]muchamadkurniawan@itats.ac.id; [3]oktavianpilu@gmail.com
*corresponding author

ABSTRACT

K-Means is the most popular of clustering method, but its drawback is sensitivity to outliers. This paper discusses the addition of the outlier removal method to the K-Means method to improve the performance of clustering. The outlier removal method was added to the Local Outlier Factor (LOF). LOF is the representative outlier's detection algorithm based on density. In this research, the method is called LOF K-Means. The first applying clustering by using the K-Means method on hotspot data and then finding outliers using the LOF method. The object detected outliers are then removed. Then new centroid for each group is obtained using the K-Means method again. This dataset was taken from the FIRM are provided by the National Aeronautics and Space Administration (NASA). Clustering was done by varying the number of clusters (k = 10, 15, 20, 25, 30, 35, 40, 45 and 50) with cluster optimal is k = 20. The result based on the value of Sum of Squared Error (SSE) shown the LOF K-Means method was better than the K-Means method.

Keywords: K-Means, LOF method, Clustering, Hotspot

Article History		
Received : June, 02nd 2020	Revised : June, 23rd 2020	Accepted : June, 30th 2020

I. INTRODUCTION

Clustering is one of the most classical algorithms of data mining methods. This algorithm is broadly utilized in numerous areas[1]. Clustering analysis can help distribute high school teachers[2]. Clustering analysis is also used in the field of e-commerce [3][4][5]. Clustering analysis's results are not only for developing e-commerce websites but also can help determine the good marketing strategy [5]. In the mining sector, clustering also can used to sort out potential areas of mining material [6]. In the field of disaster management, clustering analysis of hotspot data is performed as an effort to prevent potential forest and land fires [7],[8].

K-Means is the most popular of clustering method and is often used today [9]. K-Means has very efficient and strong elasticity compilation dealing with big data [3]. However, K-Means has a sensitivity to outliers, and this is a drawback [10]. The LOF (Local Outlier Factor) method is one of the outlier removal methods [11]. The representative outliers detection algorithm based on density is LOF [12]. Generally, the method of detecting outliers the nearest neighbor measures the absence of outliers in the context of distance in other data in the data set.

This approach risks losing outliers in the data set, where local density varies greatly. The LOF method overcomes this disadvantage by considering differences in local density around as the outlier can be measured [13]. LOF method is added to the K-Means method for improved performance of K-Means clustering. And this method is called LOF K-Means. In this paper, the results of clustering from the K-Means method and the LOF K-Means method are compared based on the Sum of Squared Error (SSE) value.

Both methods were used for clustering hotspot data and determining centroids. Hotspot data is taken from https://firms.modaps.eosdis.nasa.gov/active_fire. Why were hotspot data used in this research? Detection and analysis of hotspot data are very important to be recognized to avoid forest and land fires [7]. The results of this paper are used for our future research studies related to optimizing the suppression of forest and land fires to minimize losses.

II. METHOD

A. Hotspot Data

A hotspot is a geothermal point indicated as a fire location forest and land. In this paper, the hotspot data utilized was taken from https://firms.modaps.eosdis.nasa.gov/active_fire are provided by the National Aeronautics and Space Administration. This dataset was taken in the Southeast Asia region for seven days consecutive. The initial dataset obtained consisting of 11 features, including spatial, non-spatial, and temporal data. Then two features of spatial data are reduced, latitude, and longitude. Clustering is based on these two features.

B. K-Means

The most popular of clustering method is K-Means[9]. Dividing n object into k number of clusters so that to obtain minimum inter-cluster similarity and maximum intra-cluster similarity is the main purpose of this algorithm. Algorithm of K-Means [14] is as follow:

Input: k = number of cluster
D = dataset = $\{d_1, d_2, d_3, \ldots, d_n\}$
Method: 1. Select k point in dataset D as the beginning centroid.
2. The distance between each data point x_j and centroid was calculated. In this paper, the distance was calculated using Euclidean distance. Equation of Euclidean distance between x_j and centroid c_j, based on equation (1).

$$d(x_j, c_j) = \sqrt{\sum_{j=1}^{n}(x_j - c_j)^2} \quad (1)$$

3. Set data point into the centroid, whose distance of data point with centroid is the nearest of all centroids.
4. Recalculate the centroid k position if all the objects are placed.
5. Repeat steps 2 and 3 until the centroid k position does not change.
Output: A set of k cluster

C. LOF (Local Outlier Factor)

Flowchart determines the LOF value shown in Fig.1. LOF is comparing the local density of an object's environment with the neighboring local density based on equation (2), and (3). An object that has LOF >> 1 is called outlier. While, if an object has LOF << 1, the object is not an outlier. A high LOF value indicates that the object has a low density of its environment [11]. The LOF of $p(x_i, y_i)$ is defined as [12][15]:

$$LOF_k(p) = \frac{\sum_{o \in N_{k-dist(p)}} \frac{lrd(o)}{lrd(p)}}{|N_{k-dist(p)}|} \quad (2)$$

$$lrd(p) = 1/\left[\frac{\sum_{o \in N_{k-dist(p)}} reach-dist_k(p,o)}{|N_{k-dist(p)}|}\right] \quad (3)$$

where, $lrd(p)$ variable is local reachability density of an object p, $reach-dist_k(p,o)$ variable is reachability distance of an object p with object o, and $N_{k-dist(p)}$ variable a number of neighbors p whose distance from p is not greater than k-distance.

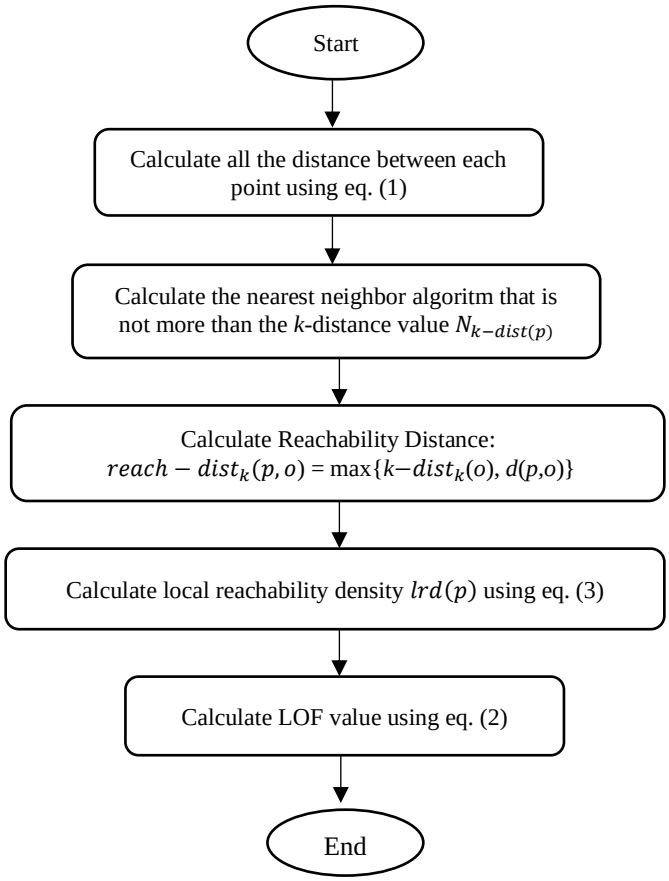

Fig.1. Local Outlier Factor Algorithm Flow Chart

DOI: 10.25139/ijair.v2i1.2634

Fig.2 illustrates the distance range with k = 4. An object p is far from o, exemplified by p_2, the distance between all is the original distance. But, if they are "close enough", the case within the figure is p_1, the original distance is supplanted by k-distance o. The statistical fluctuations of d (p,o) for all the p that are near to the variable o can be significantly reduced, that's reason for that. The parameter k can control strength of this smoothing [15].

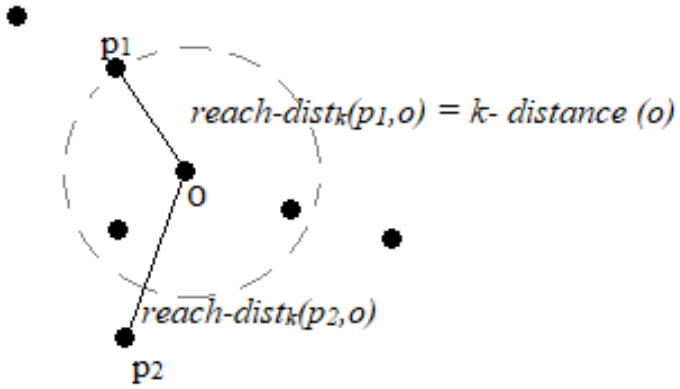

Fig.2. $reach - dist_k(p_1, o)$ and $reach - dist_k(p_2, o)$, for k=4[15]

Where, k-distance of p is defined as distance $d(p,o)$ between p and an object o is (i) for at least k objects $o' \in D\backslash\{p\}$ it holds that $d(p,o') \leq d(p,o)$, and (ii) for at most k-1 objects $o' \in D\backslash\{p\}$ it holds that $d(p,o') < d(p,o)$. In this paper $d(p,o)$ uses Euclidean distance,

$reach - dist_k(p, o) = \max\{k\text{-}dist_k(o), d(p,o)\}$, and $N_{k-dist(p)} = \{ q \in D\backslash\{p\} \mid d(p, q) \leq k\text{-distance (p)} \}$.

D. LOF K-Means

LOF K-Means is the addition of the LOF method to eliminate outliers in the K-Means clustering method. LOF K-Means description to determine the centroid of the hotspot data is shown in Fig. 3. Hotspot data is initially grouped by the K-Means method. The next step is to detect outliers for each group resulting from clustering with LOF. The object discovered outliers are then removed. Then new centroid for each group is obtained using the K-Means method again. Overall the system for clustering of hotspot data using LOF K-Means is shown in Fig.4.

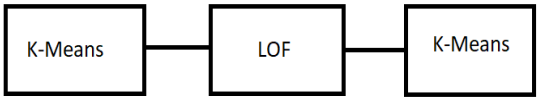

Fig.3. Block Diagram LOF K-Means

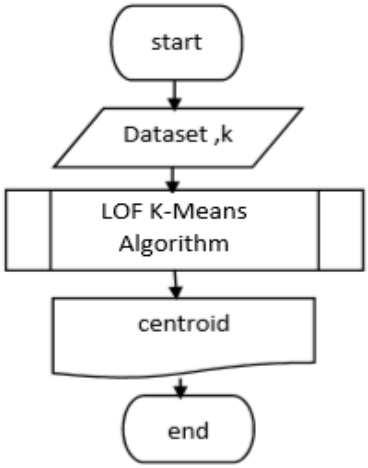

Fig.4. Clustering of Hotspot Data Using LOF K-Means Flow Chart

DOI: 10.25139/ijair.v2i1.2634

III. RESULT AND DISCUSSION

The centroid points of the hotspot data clustering using LOF K-Means results for *k=10,* and the number of outliers from each cluster for *k*=10 also shown in the table I.

TABLE I
RESULT CLUSTERING LOF K-MEANS FOR *K*=10

cluster	Non-outlier data	outlier	centroid	
			latitude	longitude
1	606	64	-2.5611384488448845	113.65633316831672
2	660	72	27.045356969696964	117.40351893939406
3	1534	262	17.181713298565818	98.3366102998696
4	465	108	-2.1122307526881743	124.12955397849468
5	559	69	22.913002862254043	109.12928658318428
6	378	57	24.61175899470897	91.93573994708998
7	773	91	25.351789521345424	102.02087490297528
8	188	21	3.6933164893617008	103.64177340425532
9	1615	343	-8.696792012383908	142.20345164086683
10	981	132	13.719616615698254	105.46438277268103

The results of the Sum of Squared Error (SSE) of both clustering methods (K-Means and LOF K-Means) are shown in Fig.5. Clustering was done by varying the number of clusters (*k* = 10, 15, 20, 25, 30, 35, 40, 45 and 50). From Figure 3, both methods have the same pattern. The optimal cluster in both methods is shown in the same cluster, cluster *k* = 20. This is shown from the biggest decrease in SSE value in cluster 20.

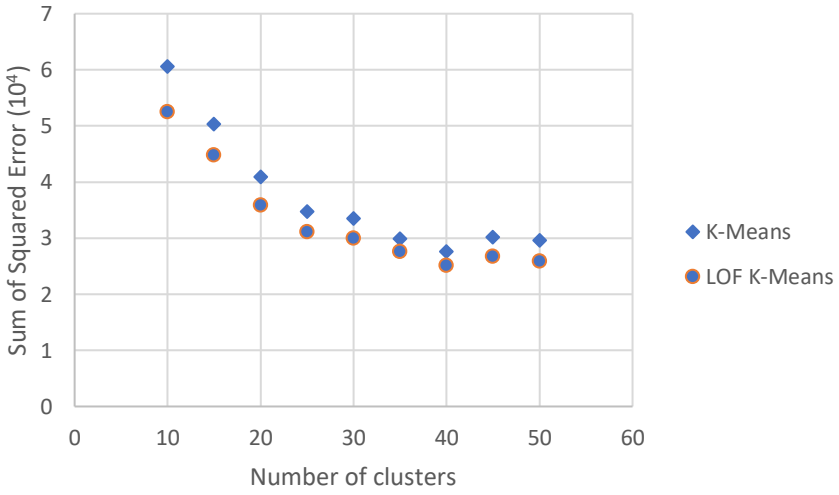

Fig.5. Sum Square Error for clustering methods K-means and LOF K-Means

The SSE value of LOF K-Means for each number of clusters is lower than the SSE value of K-Means in Fig.5. This shows the LOF K-Means method is better than the K-Means method. The average SSE value of LOF K-Means is 33285,56, while the average SSE value of K-Means is 37469,22. The best SSE value of LOF K-Means is 25147 at *k* = 40.

IV. CONCLUSION

This paper presented the clustering method for clustering hotspot data. Clustering was done by varying the number of cluster *k* = 10, 15, 20, 25, 30, 35, 40, 45 and 50. Clustering method: K-Means and LOF K-Means were evaluated for their SSE values. The evaluation results have shown that LOF K-Means was better than K-Means. Further studies are needed to be related to the outlier removal method other than LOF to be combined with the K-Means method, to obtain a better clustering method.

REFERENCES

[1] H. Z. Geng Zhang , Chengchang Zhang, “Improved K-means Algorithm Based on Density Canopy,” *Knowledge-Based Syst.*, vol. 145, pp. 289–297, 2018.

[2] T. Widiyaningtyas, M. I. W. Prabowo, and M. A. M. Pratama, "Implementation of K-means Clustering Method to Distribution of High School Teachers," *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, pp. 49–54, 2017.
[3] X. Huang and Z. Song, "Clustering Analysis on E-commerce Transaction Based on K-means Clustering," *J. Networks*, vol. 9, pp. 443–450, 2014.
[4] I. Shaik, S. S. Nittela, T. Hirwarkar, and S. Nalla, "K-means Clustering Algorithm Based on E-Commerce Big Data," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 11, pp. 1910–1914, 2019.
[5] S. Chelcea *et al.*, "Pre-Processing and Clustering Complex Data in E-Commerce Domain," in *1st Int'l Workshop on Mining Complex Data MCD'2005*, 2005.
[6] M. Lutfi, E. Sukiyah, and N. Sulaksana, "Analisis Zonasi Lahan Usaha Tambang Menggunakan Metode K-means Clustering Berbasis Sistem Informasi Geografi," *J. Teknol. Miner. dan Batubara*, vol. 15, pp. 49–61, 2019.
[7] N. L. Febriana and I. S. Sitanggang, "Outlier Detection on Hotspot Data in Riau Province using OPTICS Algorithm," in *IOP Conference Series: Earth and Environmental Science*, 2017, pp. 1–7.
[8] D. F. Pramesti, M. Tanzil Furqon, and C. Dewi, "Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 9, pp. 723–732, 2017.
[9] A. Beer, J. Lauterbach, and T. Seidl, "MORe++: k-Means Based Outlier Removal on High-Dimensional Data," in *Similarity Search and Applications 12th International Conference,SISAP*, G. Amato and C. Gennaro, Eds. Springer, 2019, pp. 188–202.
[10] P. O. Olukanmi and B. Twala, "Sensitivity analysis of an outlier-aware k-means clustering algorithm," *IEEE Pattern Recognit. Assoc. South Africa Robot. Mechatronics Int. Conf.*, pp. 68–73, 2017.
[11] N. Idham, "Penerapan Outlier Analysis sebagai Salah Satu Rekomendasi Kelompok Belajar Terhadap Siswa Kelas 6 di SDN Pagelaran II," *J. Ilm. Komput. dan Inform.*, 2017.
[12] Z. Shaomin, L. Xiangyu, and W. Baoyi, "An improved outlier delection algorithm K-LOF based on density," *Comput. Perform. Commun. Syst.*, vol. 2, no. 1, pp. 1–7, 2017.
[13] A. Mahendra, "Pentapisan dan Deteksi Data Outlier dalam Proses Sistem Akusisi Data Pada Proses Sintering," *Asitron*, vol. 6, pp. 1–7, 2015.
[14] A. Barai and L. Dey, "Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering," *World J. Comput. Appl. Technol.*, vol. 5, no. 2, pp. 24–29, 2017.
[15] M. M. Breuniq, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *SIGMOD Rec. (ACM Spec. Interes. Gr. Manag. Data)*, vol. 29, no. 2, pp. 93–104, 2000.

DOI: 10.25139/ijair.v2i1.2634