International Journal of Advances in Data and Information Systems
Vol. 1, No. 1, April 2020, pp. 9-16
ISSN: 2721-3056, DOI: 10.25008/ijadis.v1i1.13

Application of K-Means Clustering Algorithm for Determination of Fire-Prone Areas Utilizing Hotspots in West Kalimantan Province

Nabila Amalia Khairani[1], Edi Sutoyo[2]
[1,2]Department of Information System, School of Industrial and System Engineering, Telkom University

Article Info

Article history:

Received Feb 20, 2020
Revised Mar 13, 2020
Accepted Mar 26, 2020

Keywords:

Clustering
Hotspots
K-Means
Unsupervised Learning
Data mining

ABSTRACT

Forest and land fires are disasters that often occur in Indonesia. In 2007, 2012 and 2015 forest fires that occurred in Sumatra and Kalimantan attracted global attention because they brought smog pollution to neighboring countries. One of the regions that has the highest fire hotspots is West Kalimantan Province. Forest and land fires have an impact on health, especially on the communities around the scene, as well as on the economic and social aspects. This must be overcome, one of them is by knowing the location of the area of fire and can analyze the causes of forest and land fires. With the impact caused by forest and land fires, the purpose of this study is to apply the clustering method using the k-means algorithm to be able to determine the hotspot prone areas in West Kalimantan Province. And evaluate the results of the cluster that has been obtained from the clustering method using the k-means algorithm. Data mining is a suitable method to be able to find out information on hotspot areas. The data mining method used is clustering because this method can process hotspot data into information that can inform areas prone to hotspots. This clustering uses k-means algorithm which is grouping data based on similar characteristics. The hotspots data obtained are grouped into 3 clusters with the results obtained for cluster 0 as many as 284 hotspots including hazardous areas, 215 hotspots including non-prone areas and 129 points that belong to very vulnerable areas. Then the clustering results were evaluated using the Davies-Bouldin Index (DBI) method with a value of 3.112 which indicates that the clustering results of 3 clusters were not optimal.

Corresponding Author:

Edi Sutoyo,
Department of Information Systems,
Telkom University,
Bandung, West Java, Indonesia, 40257
Email: edisutoyo@telkomuniversity.ac.id

1. INTRODUCTION

Indonesia is the country that has the most extensive tropical forests and has the highest biodiversity in the world besides Brazil and Colombia. It has 120.6 million hectares or around 63 percent of the land area which is forest area. Forest fires in Indonesia attracted global attention at the time of the fires in 1982/83 and 1997/1998. In 2007, 2012 and 2015 the occurrence of forest fires in Indonesia brought cross-border haze pollution to the ASEAN region and this also became the center of attention for the global community.

Forest fires in Indonesia are already attracting the attention of the global community that brings smog pollution across borders of ASEAN. Based on information from the Bureau of Meteorology, Climatology and Geophysics in 2018, it was stated that the territory of Indonesia that was most vulnerable to forest and land fires was 11 Provinces. One area that faces the threat of a forest fire disaster is West Kalimantan Province. The impact of forest fires is very bad for the environment around the scene. From the impact caused by the forest and land fires, information is needed that can identify areas prone to hotspots. The results of this information can be used as prevention in forest fires early on [1].

Considering the factors and impacts of forest and land fires, it is necessary to be able to know the areas that will be detected in the category of areas that are prone to hotspots, areas that are very prone to hotspots and areas that are not prone to hotspots. Detection of forest and land fires is detected by satellites using the MODIS sensor on Terra / Aqua satellites and Snpp satellite VIIRS which will give an overview of hotspots in the area of forest and land fires [2]. Information from the hot spot distribution data can be seen with the highest temperature in each area of forest and land fires. Data obtained from the National Aeronautics and Space Institute (LAPAN) will be very useful information because the data can be processed using data mining techniques, where data mining techniques can process data at a large scale [3].

One study of data mining is about clustering. In the clustering algorithm, the data will be grouped into clusters based on the similarity of one data to another. The principle of clustering is to maximize the similarity between members of one cluster and minimize the similarity between members of different clusters [4].

In this research, the algorithm that will be used in the clustering method is the K-means algorithm which is one of the non-hierarchical clustering methods. The K-means algorithm will partition the data into one or more clusters. Data are grouped based on the same characteristics, but if the data has different characteristics will be grouped into other clusters / groups [5]. This method will determine the hotspot-prone areas in West Kalimantan Province using the k-means clustering algorithm. The results of the clustering data will be calculated by the Davies-Bouldin Index (DBI) method to see the number of clusters and assess whether the clustering results are optimal or not [5]. The Davies-Bouldin Index (DBI) value obtained cannot be negative and the lowest value will indicate optimal clustering.The results of this clustering are expected to be able to provide information and learning to the community in order to prevent forest and land fires in the province of West Kalimantan which includes hotspots, hotspots and hotspots. This study aims to open up the people's knowledge in Indonesia that the dangers of the impact of forest and land fires.

2. LITERATURE REVIEW

2.1. Forest Fires

In 2015, forest and land fires occurred in Indonesia which threatened Sumatra, Kalimantan and Papua. This makes 80% of Sumatra and Kalimantan covered by heavy smoke, and the incident also affects neighboring countries. As a result of the catastrophic forest and land fires, an area of 2.61 million ha of forest and land was burned with huge losses [6].

Theoretically forest fires occur due to interactions between fuel, oxygen, and heat under certain conditions. If all three elements are present together, the fire will occur. Therefore, to overcome this impact, one element was decided by removing fuel or heat [7].

2.2. Hotspots

Hotspots are by definition defined as areas that have a relatively higher surface temperature compared to the surrounding area. Monitored by remote sensing satellites based on certain temperature thresholds. Hotspots are the results of detection of forest or land fires at certain pixel sizes. The results of these fires are likely to be detected when the satellite passes under cloud-free conditions using certain algorithms. The following below is an illustration of how remote sensing satellites monitor forest or land fires in an area [8].

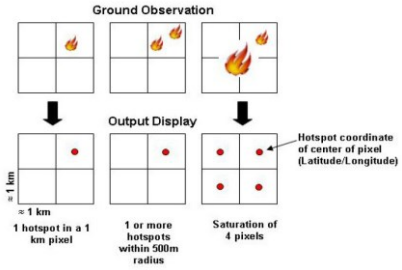

Figure 1. Illustration of hotspots

Figure 1 explains how satellites can detect fires at a location. On the left side of the satellite can detect a hotspot. The center of the satellite detects two fires within a 500m radius can be detected by only one hotspot. The right side of a very large fire can be detected 4 or more hotspots. This shows that the host point obtained is not the same as the number of events in forest and land fires in the field.

In determining the class at each point has a confidence interval or confidance level which indicates that a hotspot that is monitored via satellite by remote sensing is a true fire event that occurs in the field. If the higher the confidence interval produced, the higher the potential for the hotspot to state if there is a true forest and land fire. The MODIS Active Fire Product User's Guide divides confidence levels into three classes. The following in Table 1 below is a hotspot trust interval [9].

Table 1. Interval of confidence in hotspot information

Confidence	Class	Action
0% ≤ C < 30%	Low	Important to note
30% ≤ C < 80%	Normal	Warning
80% ≤ C ≤ 100%	High	Immediate countermeasure

2.3. Data Mining

Data mining is the process of being able to determine interesting patterns from large amounts of data, data sources in the form of databases, data warehouses, the web, other information from repositories or data that is dynamically transmitted to the system [4]. Data mining can dig up data into really data that can be used by extracting patterns from processes that cannot be known manually or from piles of data that aim to be able to make valuable information.

Analysis of data mining looks at large and complex amounts of data so that it can be used to determine conclusions and get decisions that are worth using. There are several other names of data mining, namely: knowledge discovery (mining) in databases (KDD), knowledge extraction (data extraction), data / pattern analysis, business intelligence (business intelligence), and others. The following in Figure 2 is the process of data mining.

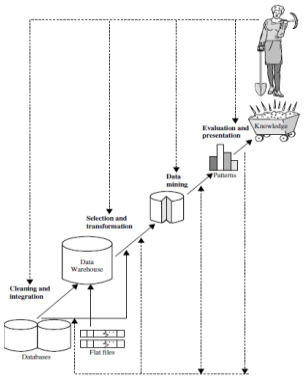

Figure 2. Data Mining Process in Knowledge Discovery in Database [4]

In general, data mining tasks can be divided into two, namely predictive and descriptive. Predictive means that data mining is done to form a knowledge model that will be used to make

predictions, such as research conducted relating to classification and prediction, forecasting, decision modeling, etc [10]–[16]. While descriptive means that data mining is done to look for patterns that can be understood by humans that explain the characteristics of the data. Some research related to descriptive such as clustering, Mining Frequent Patterns, Associations, Correlations, etc [17]–[19].

2.4. Clustering

Clustering is a data mining technique that is used to find or group data that has similar characters between one data and another. It is an unsupervised data mining method, where this method does not require training data and data testing and does not require target output [4].

Clustering consists of 2 types, namely Hierarchical and Non-Hierarchical. The difference between the two types of clustering is that Hierarchical groups two or more data by looking at the object that has the closest resemblance. Furthermore, the process is continued by looking at the closeness of the other objects until they are finished, so that the cluster will form a tree (hierarchical) level clearly between objects, starting from the most similar to the least similar. Clustering for Non-Hierarchical begins by determining the number of clusters as needed. Non-Hierarchical process is different from Hierarchical process, after the number of clusters is determined, the cluster process is directly carried out. This method is called the k-means Clustering method. In the picture below, see the difference from clustering in hierarchical and non-hierarchical clustering.

2.5. K-Means

The k-means algorithm is a simple iteration method to partition a given dataset into a number of user-defined clusters [20]. K-means is a non-hierarchical clustering method that divides data into one or more clusters. Data in one group has the same characteristics but has different characteristics from other groups. The k-means method is a distance-based clustering method that will divide data into a specified number of clusters, but this algorithm only works on numeric attributes. The use of k-means algorithm to do the clustering process basically depends on the existing data and the conclusions to be. So, for the use of the k-means algorithm in it make the following rules:

1. The number of clusters must be input.
2. Can only be done with numeric attribute types.

The k-means method will partition the data into clusters of data that have the same characteristics that will be grouped into one same cluster, then for data that has different characteristics will be grouped into other groups. The purpose of data clustering is to minimize objective fuction set in the clustering process, which generally seeks to minimize variation within a cluster and minimize variation between clusters. Following are the steps of the k-means algorithm [21].

1. Determine the number of clusters.
2. Determine the centroid point k (cluster center) randomly.
3. Calculate the distance of each point to the center of the cluster, the distance between one data with one cluster will determine which data goes into which cluster. Calculation of the distance between data and cluster center using Euclidian Distance with the formula:

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

where

$D(i,j)$: Distance of data to i to cluster center j
X_{ki} : Data to i in data attribute to k
X_{kj} : The center point to j in the data attribute to k

3. RESEARCH METHOD

The data to be used are hotspots data obtained from the National Aeronautics and Space Agency (LAPAN). The data that has been obtained is then a process, the process in data processing consists of 2 parts, namely the data cleaning process and the data clustering process.

In this process after obtaining data obtained from the National Aeronautics and Space Agency (LAPAN), the data is then cleaned and selected data based on the attributes needed for the data clustering process. Furthermore, transformation is performed on the data that has been done in the attribute selection process, where the data obtained and from the selected attributes are of type data string and to be able to do the clustering process, transformation is required on the attributes of the data type type because clustering using the k-means algorithm can only process data of numeric data type. Then it goes into the normalization stage, where the goal of the normalization stage is to range for each attribute starting from 0-1 to facilitate the clustering process and the numbers used are balanced.

In this process after the data cleaning process, the following stages in the clustering method use the k-means algorithm:

1. Determine the number of clusers
2. Determine the centroid point (cluster center) randomly.
3. Calculate the distance at each point to the center of the cluster.
4. Grouping data based on its cluster.
5. If the centroid point or center value on the cluster is still changing, a new cluster center is determined again by calculating the distance of the point to the center of the cluster.
6. Furthermore, the iteration results obtained when the center value in the cluster has not changed.

Then the final results obtained from the data clustering process. The process of implementing data clustering uses the python programming language.

In the final process an evaluation of the results of the data clustering process that has been obtained is then tested using the Davies-Bouldin Index (DBI) calculation. If the DBI value obtained is smaller and not negative, the better the cluster that results from the k-means used. The stages in the calculation of DBI are as follows:

1. Data used are the results of data processed using the euclidean distance formula that has been clustered.
2. Perform average and variance calculations for each cluster, using the formula:

$$var(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

3. Finding the value of Rij and Ri, with the formula:

$$R_i = \underset{j=1,..k, i \neq j}{max} R_{ij}$$

4. Calculating the DBI value to find out whether the results from clustering are optimal or not, using the formula:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} R_i$$

After the DBI calculation is performed, it displays a visualization of the results from the clustering data obtained. Then provide conclusions and suggestions for the clustering process with the k-means algorithm.

4. RESULT AND DISCUSSION

At this stage the K-means algorithm is implemented with the number of clusters which is 3, where this process will group data into 3 clusters. The following below are the results of data clustering with the K-Means algorithm.

4.1. Evaluation of Results

At this stage the results are evaluated to determine the optimal value in the clustering process by using the Davies-Bouldin Index (DBI) method. The results of the evaluation are as follows:

1. Calculate the average variance of data in each cluster, with the results obtained can be seen in Table 2 below.

Table 2. Average results and variance of each cluster

	Centroid 0	Centoid 1	Centroid 2
x	0.346	0.381	0.349
var	0.022	0.021	0.002

2. Calculating the R_{ij} value, with the results obtained can be seen in table 3 below:

Table 3. Results value Rij

R01	1.250
R02	8.087
R12	0.744

3. Calculating the value of R_i and the value of DBI with the results obtained can be seen in Table 4 below:

Table 4. Results of R_i values and DBI values

R1	8.087
R2	1.250
DBI	3.112

The resulting DBI value is greater than 0, so the results of clustering with 3 clusters are not optimal.

4.2. Visualization of the Results

This stage is the stage of data visualization that has been evaluated, the results of which will be implemented into a map. The following in Figure 3 is a visualization of the results implemented:

Figure 3. Location of clusters by category

In Figure 3 above you can see the visualization of the results in each cluster¸ where the red color indicates areas that are highly prone to hotspots, yellow areas that are prone to hotspots and green areas that are not prone to hotspots.

In the results of clusters for areas that are not prone to hotspots, there are not many scattered areas, for areas that are prone to hotspots, fire points can be seen in several areas and there are clustered ones, whereas for areas that are very prone to hotspots, they can be seen in several regions but not in groups. From the visualization of the results obtained, areas that are very vulnerable to hotspots are more scattered in several areas so that it can make areas that are not prone to / vulnerable to hotspots that can cause fires that have a serious impact on the surrounding community. This information can be used as a benchmark for the community to be alert to forest and land fire disasters in West Kalimantan Province.

5. CONCLUSION

Based on the results of this study, it can be concluded:

1. The k-means algorithm clustering method can be used to classify data based on categories, i.e. hotspot-prone areas, non-hotspot-prone areas and hotspot-prone areas. In 2019 data from January to June, 284 points were included which were in vulnerable areas, 215 points that were included in areas that were not vulnerable and 129 points that were classified as very vulnerable areas.
2. In the visualization of results based on clusters obtained for vulnerable areas and areas that are not prone to hotspots have the position of points that are close together so that it can cause a fire. However, for areas that are very vulnerable to hotspots, the position of points is not only in one area, but the position of the points are scattered in various regions.

REFERENCES

[1] KLHK, *Status Hutan dan Kehutanan Indonesia 2018*. 2018.
[2] K. dan G. Badan Meteorologi, "Satelit Hotspot MODIS [Indonesia Barat] BMKG," 2019. .
[3] I. Ibáñez, J. A. Silander, J. M. Allen, S. A. Treanor, and A. Wilson, "Identifying hotspots for plant invasions and forecasting focal points of further spread," *J. Appl. Ecol.*, vol. 46, no. 6, pp. 1219–1228, Nov. 2009.
[4] J. Han, M. Kamber, and J. Pei, *Data Mining, Concepts and Techniques*. 2012.
[5] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 28, no. 3, pp. 301–315, 1998.
[6] S. H. Endrawati, "Analisis Data Titik Panas (Hotspot) dan Areal Kebakaran Hutan dan Lahan tahun 2016," in *Kementerian Lingkungan Hidup dan Kehutanan*, 2016, p. 1.
[7] "Forest Fire Management," pp. 527–583, Jan. 2001.
[8] T. Amit Garg, "Modelling Fire Hazard in Pine Zone of Uttarakhand," 2017.
[9] L. Giglio, "MODIS Collection 4 Active Fire Product User ' s Guide Version 2 . 3," *Sites J. 20Th Century Contemp. French Stud.*, vol. Version 2., no. February, p. 44, 2007.
[10] E. Sutoyo, R. R. Saedudin, I. T. R. Yanto, and A. Apriani, "Application of adaptive neuro-fuzzy inference system and chicken swarm optimization for classifying river water quality," in *Proceeding - 2017 5th International Conference on Electrical, Electronics and Information Engineering: Smart Innovations for Bridging Future Technologies, ICEEIE 2017*, 2018, vol. 2018-Janua, pp. 118–122.
[11] E. Sutoyo and A. Almaarif, "Educational Data Mining untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritme Naïve Bayes Classifier," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 95–101, Feb. 2020.
[12] I. T. R. Yanto, E. Sutoyo, A. Apriani, and O. Verdiansyah, "Fuzzy Soft Set for Rock Igneous Clasification," in *2018 International Symposium on Advanced Intelligent Informatics (SAIN)*, 2018, pp. 199–203.
[13] A. P. Slavia, E. Sutoyo, and D. Witarsyah, "Hotspots Forecasting Using Autoregressive Integrated Moving Average (ARIMA) for Detecting Forest Fires," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 2019, pp. 92–97.
[14] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybern. Inf. Technol.*, vol. 13, no. 1, pp. 61–72, 2013.

[15] A. Aninditya, M. A. Hasibuan, and E. Sutoyo, "Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 2019, pp. 112–117.
[16] E. Sutoyo and A. Almaarif, "Twitter Sentiment Analysis of The Relocation of Indonesia's Capital City," *Bull. Electr. Eng. Informatics*, vol. 9, no. 04, 2020.
[17] E. Sutoyo, I. T. R. Yanto, Y. Saadi, H. Chiroma, S. Hamid, and T. Herawan, "A Framework for Clustering of Web Users Transaction Based on Soft Set Theory," in *Proceedings of the International Conference on Data Engineering 2015 (DaEng-2015)*, 2019, vol. 520, pp. 307–314.
[18] E. Sutoyo, I. T. R. Yanto, R. R. Saedudin, and T. Herawan, "A soft set-based co-occurrence for clustering web user transactions," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 15, no. 3, 2017.
[19] T. F. Gharib, H. Nassar, M. Taha, and A. Abraham, "An efficient algorithm for incremental mining of temporal association rules," *Data Knowl. Eng.*, vol. 69, no. 8, pp. 800–815, Aug. 2010.
[20] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.
[21] J. Wu, *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media, 2012.