Web-Based Application for Outliers Detection on Hotspot Data Using K-Means Algorithm and Shiny Framework

Agisha Mutiara Yoga Asmarani Suci1, Imas Sukaesih Sitanggang2

Computer Science Department, Faculty of Mathematics and Science
Bogor Agricultural University, Bogor 16680, Indonesia

agishamyas@gmail.com1, imas.sitanggang@ipb.id2

Abstract. Outliers analysis on hotspot data as an indicator of fire occurrences in Riau Province between 2001 and 2012 have been done, but it was less helpful in fire prevention efforts. This is because the results can only be used by certain people and can not be easily and quickly accessed by users. The purpose of this research is to create a web-based application to detect outliers on Hotspot data and to visualize the outliers based on the time and location. Outliers detection was done in the previous research using the k-means clustering method with global and collective outlier approach in Riau Province Hotspot data between 2001 and 2012. This work aims to develop a web-based application using the framework Shiny with the R programming language. This application provides several functions including summary and visualization of the selected data, clustering hotspot data using k-means algorithm, visualization of the clustering results and sum square error (SSE), and displaying global and collective outliers and visualization of outlier spread on Riau Province Map.

1. Introduction
In early March 2014, forest and peatland fires in Riau province soared to beyond the current crisis haze in Southeast Asia in June 2013. Since 20 February to March 11th, 2014, Global Forest Watch 3101 detects fires warning with a high confidence level in Sumatra using active hotspot data from NASA [1]. These numbers exceeds the 2643 total number of hotspots detected warnings at the peak of the crisis fires and haze on 13-30 June 2013 [1]. NASA active hotspot data obtained from satellite observations that use the sensor called the Moderate Resolution Imaging Spectroradiometer (MODIS).

Strategies that can be done as an effort to prevent the occurrence of such fires is the fire information system approach to system monitoring hotspots [2]. The hotspots data can be used as one indicator of forest and land fires events, so that the necessary analysis. Analysis can be done to the incident hotspots, namely the analysis of outliers using clustering algorithm [2].

Previous research about analysis outliers on hotspot data have been carried out [3] using the K-Means algorithm. However, in these studies only use the functions contained in R and not created a Web-based system. Results of the research that object is used as outliers is an object with a highest hotspots frequency. In other previous research, data mining technique has been applied on hotspot dataset in Riau from 2000 until 2014 using sequential pattern mining algorithm namely prefix span to find frequent sequences [4]. Clospan (Closed Sequential Pattern Mining) that available in the SPMF (Sequential Pattern Mining Framework) program to generate sequential patterns on hotspots data was...
used on hotspots, precipitation, and temperature data that are grouped by year of events starting from 2001 to 2010 [5]. Another, spatial decision tree algorithm on data of forest fires has been done. The algorithm is the improvement of the conventional decision tree algorithm in which distance and topological relationships are included to grow spatial decision trees [6]. Decision tree to predict hotspot occurrences in Bengklis district, Riau Province using the spatial entropy-based decision tree algorithm was used. Data was used by that algorithm include city centre, river, road, income source, rain cover, population, precipitation, school, temperature, and wind speed [7]. Not only decision tree, extended spatial ID3 algorithm on the forest fires data in Bengkalis district, Riau Province Indonesia used too. The data include hotspots and non-hotspots, weather data, socio-economic data, and geographical characteristic. The result is a decision tree with the income source layer as the label of root node [8]. Association rule algorithm namely apriori to discovers the possible influence factors on the occurrence of fire events in the study area of Rokan Hilir Riau Province Indonesia was done. The Apriori algorithm was applied on a forest fire dataset which contained data on physical environment (land cover, river, road and city center), socio-economic (income source, population, and number of school), weather (precipitation, wind speed, and screen temperature), and peatlands [9]. ID3 and C4.5 that are decision tree algorithm was used in previous research on spatial data of forest fires for Rokan Hilir District in Riau Province in Indonesia. The result show that C4.5 algorithm has better performance than the ID3 algorithm in terms of accuracy and the number of generated rules [10]. Improvement of the ID3 decision tree algorithm such that it can be utilized on spatial data in order to develop a classification model for hotspots occurrence. As the ID3 algorithm that uses information gain in the attribute selection, the proposed algorithm uses spatial information gain to choose the best splitting layer from a set of explanatory layers. The new formula for spatial information gain is proposed using spatial measures for point, line and polygon features. The proposed algorithm has been applied on the forest fire dataset for Rokan Hilir district in Riau Province in Indonesia. The dataset contains physical data, socio-economic, weather data as well as hotspots and non-hotspots occurrence as target objects [11]. Classification module for hotspot occurrences in a GIS using OpenGeo Suite 3.0. The classification module was made using the decision tree method namely the C4.5 algorithm. The GIS enables users to determine whether an area is potential for hotspot incidence or not based on its characteristics. Potential hotspot occurrences can be used in decision making related to forest fire prevention [12].

Almost all the people of Indonesia need the internet. By using the internet makes it possible to perform long-distance communications quickly and cheaply. Results of the outlier detection of hotspots data need to be distributed quickly for stakeholders to be able to take a decision as early as possible in an effort to anticipate the fires and fire suppression.

Shiny is a package of R programming language that allows users to build web-based applications. Shiny provide queries and summary of the data to the user via a modern web browser easily. In addition, Shiny also equipped with a variety of widgets for building interactive user interface and can be easily integrated with HTML and CSS, JavaScript and JQuery can even be used to expand the scope of application Shiny [13].

2. K-Means Algorithm

K-Means algorithm is a technique based centroid and is one method of partitioning. $p \in C_i$ difference between an object and $c_i$, a representation of the cluster measured by $dist(p, c_i)$. $C_i$ is the i-th cluster, while the cluster centroid $c_i$ adalah to-i. $Dist(x,y)$ is the Euclidean distance between two objects or points x and y. The quality of a cluster $C_i$ can be measured by Sum Square Error (SSE). SSE is defined as follows [14]:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2 \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2 \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2$$

(1)

SSE is the total of the squared error for all objects in the data set, $p$ is a point that represents the object, $k$ as the number of clusters, and $c_i$ is the focal point cluster $C_i$ [6].
3. Outlier detection Clustering Based on Hotspot Data

This section describes fundamentals of outlier detection on hotspot data in Riau, including definition of outlier, detection outliers method, type of outliers, and degree of being an outliers. Outlier is a data object that significantly deviates from a collection of objects that are generated by different mechanisms [14].

Outliers detected with clustering method. Clustering is the process of grouping a set of data objects into groups or clusters so that objects within a cluster have high similarity, but very different from the other groups [14]. A cluster is a collection of records that are similar to one another and different from the other records in the group [15].

There are two various outliers will be detected in application that will be built. That outliers are global outlier and collective outlier [14]. Global outlier is a data object that significantly deviates from hotspot data or not in accordance with the pattern in general. Collective outlier is part of the data objects are entirely deviate significantly from the overall hotspot data.

In k-means clustering method, an objects came as global outlier outliers when the distance between the object to the center or centroid of cluster is large. Global outlier approach can be measured by the value of outlier score. Object has greater outlier score value is become a global outlier.

In addition, a collective approach to outliers in clustering method when the object is part of a small cluster member or cluster minority, then all objects in the cluster is an outlier [14]. Outlier detection also provide specific methods to calculating the degree of which data measurements deviate from the normal pattern of hotspot data. In hotspot data, outliers are measured in two scales, i.e., scalar and outlier score [16].

The scalar scale is a zero-one classification measure [16], which classifies each data measurement into outlier cluster. Thus, the output of techniques of scalar scale is provide a list of outlier that has a minimum member in clusters.

Techniques of scale value outliers set scores for each measurement data that depends on the degree of measurement are considered as outliers and provide a ranked list outliers. An analyst can choose to analyze the top n outliers that have the biggest outlier values or use a cut-off threshold to select outliers. This limit is often difficult to choose and are usually determined and set by the user [16].

Ratio to measure an object distance is defined as follows [14]:

\[
\text{Outlier Score} = \frac{\text{dist}(o, c_o)}{l_{c_o}}
\]

\(o\) : object of data,

\(c_o\) : centroid,

\(\text{dist}(o, c_o)\) : Euclid distance between object with the nearest centroid of object

\(l_{c_o}\) : mean \(\text{dist}(o, c_o)\).

4. Web-based Application for Outlier Detection

Application was built by using Shiny framework and R language for outlier detection on hotspot data between 2001 until 2012 in Riau Province with k-means clustering algorithm. This application consists of 5 menus, data menu, clustering menu, outlier detection menu, help menu, and about menu. 5 display menu available using NavbarPage, that is menus that are available in the application form of the navigation menu.

Users can choose which datasets to be used for the process of clustering on the data with a dropdown menu button. The dataset consists of datasets daily frequency of hot spots Riau Province between 2001 and 2012. The object of this dataset is made reactive to respond to any changes in user demand data. Data menu consists of two sub-menus, which is a data summary sub-menu and plot data sub-menu.

In the sub-menu of data summary, the data that has been selected will be displayed in the form of summaries minimum value, first quartile, median, average value, the third quartile, and the maximum value of the frequency. Display of data summary sub-menu can be seen in Figure 1.
While in the data plot sub-menu is displayed line graph that showing the trends or distribution of the frequency of hotspots on the day index. The graph on the data plot sub-menu made reactive. Reactive means responsive to changes in inputs, so that the plot will vary according to the selected data input by the user. Display sub-menu of data plot can be seen in Figure 2.

If users want to do clustering, then the user must select the menu clustering. K-means function on this application is reactive, so any changes in the value of the input will cause the redial function of the k-means and generate new cluster without the need to reload the browser page. Clustering menu consists of three sub-menus, namely clustering summary sub-menu, sum square error sub-menu, and k-means plot sub-menu.

Objects cluster, size of the cluster, the center of the cluster and the percent that is the result of clustering are combined to form a new frame of data to be displayed on the sub-menu clustering summary. Display of clustering summary sub-menu can be seen in Figure 3.

Object number of clusters, SSE, total withins and percent which is the clustering results are combined to form a new frame of data to be displayed on error square sum sub-menu. Display sub-menu sum square error can be seen in Figure 4.
In the k-means plot sub-menu, displayed visualization of the clustering results. To visualize the clustering results are used ggplot2 library. Display sub-menu K-Means plot can be seen in Figure 5.

![Figure 4. Sum square error sub-menu](image)

![Figure 5. K-means plot sub-menu](image)

Outlier detection menu is a menu that serves to detect outliers. The menu consists of two menus, that is collective and global outlier menus. If the user wants to detect collective outliers, then the user must select the collective outlier menu, whereas if you want to detect global outlier, then select a global menu.

On the menu collective outlier users can choose a minimum percentage of the number of members in a cluster by entering the percentage value in the column that has been provided. The menu consists of collective outlier summary sub-menu and collective outlier plot sub-menu.

Collective outlier summary sub-menu displays the results of the collective outlier detection in general. Detection results in the form of a table which consists of cluster objects and percent. If the user wants to see more clearly the objects which are included collective outliers, then the user can select or press the button detailed information available on the outlier collective summary sub-menu. Display collective outlier summary sub-menu can be seen in Figure 6.

![Figure 6. Collective outlier summary sub-menu](image)

Sub-menu displays visualization collective plot outlier outlier detection results collectively. Latitude and longitude coordinates of hot spots were detected as outliers collective visualized into a map of the province of Riau. Collective outliers visualized by month for hotspots which are distinguished by color. Display sub-menu collective outlier plot can be seen in Figure 7.
On the global outlier menu, user can select the desired amount based on the value outlier best scores in the column that has been provided. Outlier global menu consists of two sub-menus, they are global outlier summary sub-menu and global outlier plot sub-menu. Global outlier summary sub-menu displays the outlier detection results. Detection results in the form of a table which consists of no object, index, frequency, score, date, and clusters. Display sub-global menu outlier summary can be seen in Figure 8.

Global outlier plot sub-menu display the visualization of the results of a global outlier detection. Latitude and longitude coordinates of hotspots were detected as a global outlier visualized into maps Riau Province. Global outliers visualized by month for hotspots which are distinguished by color. Display sub-global menu outlier plot can be seen in Figure 9.

Available main functions testing on the application done using blackbox testing. After the tests blackbox, the available main function in the application is running according to its function. That
functions are data summary, data plot, clustering summary, SSE, K-means plot, Collective outlier summary, collective outlier plot, global outlier summary, gabol outlier plot, help, and about.

5. Conclusion
Web-based applications for outlier detection of hotspots using k-means clustering algorithm and Shiny framework provides several major functions. First, summary and plot of the selected data. Second, clustering hotspots data with k-means algorithm. Third, visualization of the clustering results. Fourth, show the SSE of clustering result. Fifth, global and collective outliers detection. Last, visualization the outliers on the Riau Province map. All of main function that available on the application works well.

References
[1] Sizer N, et al. Kebakaran hutan di Indonesia mencapai tingkat tertinggi sejak kondisi darurat kabut asap Juni. 2014. Available: http://www.wri.org/blog/2014/03/kebakaran-hutan-di-indonesia-mencapai-tingkat-tertinggi-sejak-kondisi-darurat-kabut (URL).
[2] Adinugroho WC, Suryadiputra INN, Saharjo BH, Siboro L. Panduan Pengendalian Kebakaran Hutan dan Lahan Gambut. Bogor (ID): Wetlands International-Indonesia Programme and Wildlife Habibat Canada, 2005.
[3] Imas SS and Dhiya AMB. Global and Collective Outliers Detection on Hotspot Data as Forest Fires Indicator in Riau Province, Indonesia. Proceedings 2015 Second IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM 2015), Fuzhou, China, Jul. 8-10, 2015, pp 66 – 70.
[4] Nida ZN. Sequential Pattern Mining on Hotspot data in Riau Province Using The PrefixSpanAlgorithm. Proceedings of 2015 Third International Conference of Adaptive and Intelligent Agroindutry (ICAIA) 2015, pp 313-316.
[5] Tria A and Imas SS. Sequential Patterns for Hotspots Occurences Based Weather Data Using Clospan Algorithm. Proceedings of 2015 Third International Conference of Adaptive and Intelligent Agroindustry (ICAIA) 2015, pp 301-305.
[6] Imas SS, Razali Y, Norwati M, Ainuddin AN. A Decision Tree Based on Spatial Relationship for Predicting Hotspots in Peatlands. TELKOMINKA, pp 12(2): 511-518.
[7] Indry DN and Imas SS. Classification Rules for Hotspots Occurences Using Spatial Entropy-based Decision Tree Algorithm. The 1st International Symposium on LAPAN-IPB Satellite for Food Security and Environmental Monitoring, Procedia Environmental Sciences, pp 24(2015): 120-126.
[8] Yaumil MK and Imas SS. A Spatial Decision Tree Based on Topological Relationships for Classifying Hotspot Occurences in Bengkalis Riau Indonesia. ICACSIS 2014, IEEE, pp 268-272.
[9] Imas SS. Spatial Multidimensional Association Rules Mining in Forest Fire Data. Journal of Data Analysis and Information Processing (JDAIP) 2013,1, pp 90-96.
[10] Imas SS, Razali Y, Norwati M, Ainuddin AN. Predictive Models for Hotspots Occurrence Using Decision Tree Algorithms and Logistic Regression. Journal of Applied Science 2013, ISSN Asian Network for Scientific Information, pp 13(2): 251-261.
[11] Imas SS, Razali Y, Norwati M, Ainuddin AN. Classification Model for Hotspot Occurrences Using Spatial Decision Tree Algorithm. Journal of Computer Science 2013, ISSN Science Publications, pp 9(2): 244-251.
[12] Khairil A and Imas SS. A Geographical Information System for Hotspot Occurrences Classification in Riau Province Indonesia . The 1st International Symposium on LAPAN-IPB Satellite for Food Security and Environmental Monitoring, Procedia Environmental Sciences, pp 24(2015): 127-131.
[13] Beeley C. Web Application Development with R Using Shiny. Birmingham (UK): Packt, 2006.
[14] Han J, Kamber M, and Pei J. Data Mining: Concepts and Techniques 3rd ed. Massachusetts. US: Morgan Kaufman, 2012.
[15] Larose DT. Discovering Knowledge in Data: An Introduction to Data Mining. New Jersey (US): J Wiley, 2005.

[16] Zhang Y, Meratnia N, Havinga P. Outlier detection techniques for wireless sensor networks. IEEE Communications Surveys & Tutorials, Second Quarter 2010 pp 12(2): 1-12.