MATTER: International Journal of Science and Technology
ISSN 2454-5880

Ramadhan et al., 2017

Volume 3 Issue 2, pp. 588-597

Date of Publication: 10th November 2017

DOI-https://dx.doi.org/10.20319/mijst.2017.32.588597

This paper can be cited as: Ramadhan, M., Sitanggang, I., & Anzani, L. (2017). Classification Model for Hotspot Sequences as Indicator for Peatland Fires using Data Mining Approach. MATTER: International Journal of Science and Technology, 3(2), 588-597.

CLASSIFICATION MODEL FOR HOTSPOT SEQUENCES AS INDICATOR FOR PEATLAND FIRES USING DATA MINING APPROACH

Muhammad Murtadha Ramadhan
Affiliation with Bogor Agricultural University, Bogor, Indonesia
muhmurtadha29@gmail.com

Imas Sukaesih Sitanggang
Affiliation with Bogor Agricultural University, Bogor, Indonesia
imas.sitanggang@apps.ipb.ac.id

Larasati Puji Anzani
Affiliation with Bogor Agricultural University, Bogor, Indonesia
larasatipujianzani@gmail.com

Abstract

One action which can be taken to avoid forest and land fires is to predict where forest and land fires are likely to happen. This can be done by predicting the hotspot as one of forest fires indicators. A hotspot that appears in a sequence for 2 – 5 days can be a strong indicator of forest fires. This study aims to develop prediction model for hotspot emergence in peatlands in Sumatra in 2014 and 2015 using data mining approach. The classification algorithms used are C5.0 and Random Forest which are categorized in Decision Tree model. C5.0 additionally results a rule-based model. Accuracy of the decision tree model and the rule-based model from C5.0 and Random Forest on the dataset of 2014 is 96.8%, 96.0%, and 85.6%, respectively. Accuracy of

the decision tree model and the rule-based model from C5.0 and Random Forest on the dataset of 2015 is 97.1%, 96.6%, and 75.6%, respectively. The attributes that appear from the hotspot classification model are peatlands depth and peatlands type. Hotspots in sequence are most predicted to happen on peatland that have characteristics such as type of peatlands hemist, saprists or fibrists, peatland depth is shallow, medium or deep, and can happen in every type of land use that are used for plantation or other purposes. Field verification is required to be conducted in the future in order to evaluate the prediction model

Keywords

C5.0, Classification, Data Mining, Forest and Land Fires, Hotspots, Random Forest

1. Introduction

Forest and land fires occur frequently in Sumatra and Kalimantan islands, Indonesia. In 2015 forest and land fires are the largest event in the last decade. Those fire events especially in peatland cause many negative impacts for human life and contribute to the global warming. Haze and smoke produced from the peatland fires influenced people not only in Indonesia but also in neighboring countries including Malaysia and Singapore. This situation leads to the issue of transboundary haze. Hotspot is still used as one of indicator for forest and land fires in Indonesia. Hotspot datasets are provided by some international institutions such as FIRMS MODIS NASA (http://earthdata.nasa.gov) and local institutions namely Ministry of Forestry and Environment, Republic of Indonesia.

Hotspots are recorded every day resulted in large datasets. Studies on analyzing hotspot data and other influencing spatial data was conducted in our previous works based on data mining approach including classification and sequential pattern mining. Primajaya, Sitanggang, & Syaufina (2017) developed a visualization module for spatial decision as the model for predicting hotspot occurrences. Siknun & Sitanggang (2016) developed a web-based application for hotspot classification based on decision tree models. Nurpratami & Sitanggang (2015) applied the spatial entropy based decision tree algorithm for hotspot occurrences classification. Khoiriyah & Sitanggang (2014) constructed a spatial decision tree based on topological relationship for classifying hotspot occurrences in Bengkalis Riau Indonesia. Although a hotspot is used as an indicator for forest and land fires, not all hotspots indicate real fires. However,

hotspots that are occurred consecutively in 2 to 5 days in the same area indicate strong indicator for forest fires.

Our previous studies have successfully identified sequence patterns on hotspot datasets in Kalimantan and Sumatra. Abriantini et al. (2017) identified sequences of hotspot based on sequential pattern mining approach. Sitanggang & Fatayati (2016) generated sequential pattern of hotspot occurrences in order to identify fire spot in Sumatra Island Indonesia in 2014 and 2015. Nurulhaq & Sitanggang (2015) used the PrefixSpan algorithm to generate sequence pattern on hotspot dataset in Riau Province. Agustina, & Sitanggang (2015) resulted sequence patterns of hotspot based on weather data.

This study applied performed classification tasks using decision tree algorithm on sequential pattern of hotspots. The main objective of this study is to determine characteristics of hotspot sequences and to create classification model for identifying hotspots in sequences. The classification model for hotspot sequences can be used as an early warning system for forest and land fires. The paper is organized as follows: introduction is in section 1. Section 2 shows literature review regarding previous research related to this paper. Research methodology including classification algorithms is briefly discussed in section 3. Section 4 explains results and discussion. Finally we summarize the conclusion in section 5.

2. Literature Review

Sitanggang, Putri, Khotimah, & Syaufina (2017) studied characteristics of hotspot sequential patterns in peatland using data mining approach namely sequential pattern mining. The Sequential Pattern Discovery using Equivalent Class (SPADE) algorithm was implemented on hotspot data which were collected from FIRMS NASA. The results are 316 sequences found in Kalimantan and Sumatra in 2014 and 2015 with a minimal of two days occurrences and minimum support of 1%. Wijaya, Sitanggang, & Syaufina (2016) applied the clustering algorithm on hotspot data with road and river as obstacles. The algorithm used in CPO-WCC (Clustering in Presence of Obstacles with Computed number of Cells) algorithm. The results are three clusters of hotspot. The highest number of hotspots occurrence is found in peatland with type of Hemists /Saprists (60/40) and depth greater than 400cm. Sequential patterns on hotspot datasets in Sumatra Island Indonesia in 2014 and 2015 was generated using the SPADE algorithm (Sitanggang & Fatayati, 2016). The study shows that sequence patterns of hotspot in

Sumatra in 2014 were occurred in the villages which have the weather conditions: average humidity between 70.1% and 78.2%, average temperature between 26.50 ˚C and 27.89˚C, and precipitation of 0 and 0.9 mm. In addition, sequence patterns of hotspot in Sumatra in 2015 were occurred in the villages which have the weather conditions: average humidity between 68.6 and 77.7%, average temperature between 27.00 and 27.69˚C, and precipitation of 0 mm. Spatial decision tree algorithms were applied to classify classes before burned, burned and after burned from remote sensed data of peatland area in Kubu and Pasir Limau Kapas subdistrict, Rokan Hilir, Riau (Thariqa, Sitanggang, & Syaufina, 2016). The algorithms used are Classification and Regression Trees (CART), C5.0, and C4.5. The experimental results showed that the C5.0 algorithm generates the most accurate classifier with the accuracy of 99.79%. Kirana, Sitanggang, & Syaufina (2015) identified the distribution pattern of hotspot clusters in the peatland areas in Sumatera in the year 2014 using Kulldorff's Scan Statistics (KSS) method with Poisson model. Results are clusters of hotspots which have the accuracy of 95%. Riau and South Sumatera province have the highest density of cluster distributions of the hotspot. Based on the maturity level of peat, cluster distributions of hotspot were mostly found in 'hemic' maturity level. Based on peatland thickness, cluster distribution of hotspot was mostly found in 'very deep' thickness. Sitanggang, Yaakob, Mustapha, & Ainuddin, (2015) applied three decision tree algorithms i.e. ID3, C4.5 and extended spatial ID3 on a dataset containing 235 objects that have the true class and 326 objects that have the false class. The results are decision trees for modeling hotspots occurrence which have the accuracy of 49.02% for the ID3 decision tree, 65.24% for the C4.5 decision tree, and 71.66% for the extended spatial ID3 decision tree. In another hand, classification models were developed by using a spatial decision tree algorithm on spatial data of forest fires (Sitanggang, Yaakob, Mustapha, & Ainuddin, 2014). The result is a pruned spatial decision tree with 122 leaves and the accuracy of 71.66%. The spatial tree has produced higher accuracy than the non-spatial trees that were created using the ID3 and C4.5 algorithm. The ID3 decision tree had accuracy of 49.02% while the accuracy of C4.5 decision tree reached 65.24%.

3. Methods

3.1 Decision Tree Algorithm

Classification is a method to learn a target function f that maps each attribute set x to predefined class label y in which the target function is also known classification model (Tan,

Steinbach, & Kumar, 2005). The classification model used is decision tree model and rule-based model. A decision tree is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions (Han et al., 2011). Besides, a rule-based model is a set of if-then conditions that have been derived from a tree model into more simple conditions. Tree algorithm for data mining approach in this study are C5.0 and Random Forest, however one of which with the best accuracy will be selected for creating the best classification model.

3.2 Random Forest Algorithm

Random Forest is categorized as an ensemble learning method that generates many classifiers and aggregate their result for prediction (Liaw & Wiener, 2002). Breiman (2001) explains that random forests are used in classification process by combining tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. In addition, random forests have different construction with standard classification or regression trees. Liaw & Weiner (2002) explain that standard trees use the best split among variables in splitting each node, whereas each node is split using the best among a subset of predictors randomly chosen at that node.

3.3 C5.0 Algorithm

C5.0 as a method in decision tree works by testing the classfier first to classify unseen data and for this purpose resulting decision is used (Pandya & Pandya, 2015). Pandya and Pandya (2015) in their study conclusively explains that C5.0 is an improvement of C4.5 algorithm which is faster in processing time, more efficient in memory usage, lower in error, and ultimately more accurate for classification. For details, C5.0 has several features which are:

1. The large decision tree can be viewing as a set of rules which is easy to understand.
2. C5.0 gives the acknowledge on noise and missing data.
3. Problem of over fitting and error pruning is solved by the C5.0.
4. In classification technique the C5.0 classifier can anticipate which attributes are relevant and which are not relevant in classification.

4. Result and Discussion

4.1 Dataset

The dataset used in this study is hotspot in Sumatera island in 2014 and 2015. The dataset was obtained from National Aeronautics and Space Administration (NASA) Fire Information for Resource Management (FIRMS) (http://earthdata.nasa.gov). Additional data which are peatland depth and peatland type were obtained from Wetland International. Then, the dataset was pre-processed to get clean data for being classified. Besides using hotspot data with CSV format, this study also used land use data in Sumatera in the SHP format. The attributes ultimately used for classification process are peatland type, peatland depth, and land use in the study area.

The peatland types were differetianted to three types according to its decomposition which are fibric organic, hemik, and sapric (Wahyunto et al., 2005). Peatland depth can be categorized into D0 (d<50 cm), D1 (50 cm<d<100 cm), D2 (100 cm<d<200 cm), D3 (200 cm<d<400 cm), and D4 (d>400 cm) (www.wetlands.org; Sitanggang et al. 2012).

4.2 Data Preprocessing

The dataset obtained at first consisted of sequence hotspot and non-sequence hotspot in Sumatera in 2014 and 2015. The dataset still included hotspot which is located out of peatland before preprocessing. Representative hotspot data which are on the peatland were selected to be used in classification process.

Peatland data in the SHP format were integrated with hotspot data in format CSV. In integration process, coordinate reference system of both data had to be in the same on map or data projection. Buffer process was implemented on Quantum GIS for the sequence hotspot data until the hotspot data were differentiated to sequence and non-sequence with distance of 0.01 (in degree). The result of dissolve process in hotspot buffer can be seen in Figure 1.

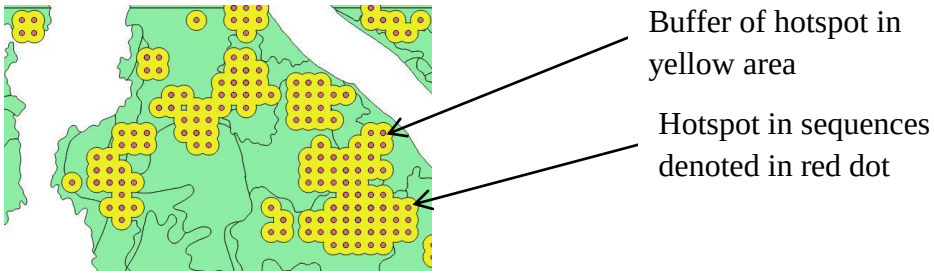

Figure 1: *Buffer and dissolve result of hotspot in sequence*

After buffer process, necessary hotspot data on peatland were selected with clip feature in Quantum GIS. Besides, hotspot data which were out of peatland (non-sequence hotspot on

peatland) were separated selected with difference feature in Quantum GIS. In the last step, the dataset was generated using database query in PostgreSQL. Data processed in Quantum GIS were sent to PostgreSQL with the spatial extension PostGIS and dataset were generated with two class values which are TRUE (sequence hotspot) and FALSE (non-sequence hotspot).

4.3 Classification with C5.0 and Random Forest

This study implemented two classification algorithms which are C5.0 and Random Forest. The accuracy as results of peatland hotspot classification in Sumatera in 2014 and 2015 with C5.0 (decision tree model rule-based model) and Random Forest are provided in Table 1. In hotspot dataset 2014, the total sample data used are 8716 and total attributes used are 3 attributes (peatland type, peatland depth, and landuse) with factor type from hotspot in peatland. We applied 10-fold cross validation in calculating accuracy of classifiers. Classification by C5.0 resulted in accuracy of 96.8% for decision tree model and 96.0% for rule-based model while Random Forest resulted in accuracy of 85.6%. The results show that tree model in C5.0 is better in classifying the dataset than Random Forest with accuracy of 96.8%.

Table 1: *Prediction result of peatland hotspot classification in Sumatera in 2014 and 2015 using decision tree model, rule-based model, and random forest model*

Prediction Results	Hotspot in Sequence		Non-Sequence Hotspot	
	2014	2015	2014	2015
Decision Tree Model (C5.0)				
Hotspot in Sequence	4674	4391	110	142
Non-Sequence Hotspot	170	117	3762	4423
Rule-based Model (C5.0)				
Hotspot in Sequence	4652	4333	132	200
Non-Sequence Hotspot	219	106	3713	4434
Random Forest Model				
Hotspot in Sequence	3304	3136	1094	1091
Non-Sequence Hotspot	160	1109	4158	3737

In hotspot dataset 2015, the total sample data used are 9073 and total attributes used are 3 attributes (peatland type, peatland depth, and landuse) with factor type from hotspot on Sumatera peatland. Classification by C5.0 resulted in accuracy of 97.1% for decision tree model and 96.6% for rule-based model while Random Forest resulted in accuracy of 75.6%. The results show that tree model in decision tree model of C5.0 is better in classifying the dataset than Random Forest with accuracy of 97.1%. The attributes that appear from the hotspot classification model are

peatlands depth and peatlands type. Hotspots in sequence are most predicted to happen on peatland that have characteristics such as type of peatlands hemist, saprists or fibrists, peatland depth is shallow, medium or deep, and can happen in every type of land use that are used for plantation or other purposes.

5. Conclusion

This study implemented two classification algorithms in data mining which are C5.0 and Random Forest in order to generate model for sequence hotspot prediction on hotspot datasets in Sumatera island in 2014 and 2015. The hotspot datasets were obtained from National Aeronautics and Space Administration (NASA) Fire Information for Resource Management (FIRMS) by request. The classification was conducted to 8716 sample data from hotspot data in Sumatra 2014 and 9073 sample data from hotspot data in Sumatra 2015. Experimental results show that decision tree model of C5.0 provides the best accuracy in prediction compared to rule-based model of C5.0 and Random Forest model. In conclusion, decision tree model of C5.0 as a data mining technique is highly representative to classify hotspot sequences as indicator for peatland fires. The limitation of this study is that the classification models for hotspot sequences have not been verified. Therefore, the field work will be done in the future to verify the models.

References

Abriantini, G., Sitanggang, I. S., & Trisminingsih, R. (2017, January). Hotspot sequential pattern visualization in peatland of Sumatera and Kalimantan using shiny framework. In *IOP Conference Series: Earth and Environmental Science* (Vol. 54, No. 1, p. 012057). IOP Publishing. https://doi.org/10.1088/1755-1315/54/1/012057

Agustina, T., & Sitanggang, I. S. (2015, August). Sequential Patterns for hotspot occurrences based weather data using Clospan algorithm. In *Adaptive and Intelligent Agroindustry (ICAIA), 2015 3rd International Conference on* (pp. 245-249). IEEE. https://doi.org/10.1109/ICAIA.2015.7506514

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Kirana, A. P., Sitanggang, I. S., & Syaufina, L. (2015). Poisson Clustering Process on Hotspot in Peatland Area using Kulldorff's Scan Statistics Method. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *13*(4), 1376-1383. https://doi.org/10.12928/telkomnika.v13i4.2272

Khaira, U., Sitanggang, I. S., & Syaufina, L. (2016). Detection and Prediction of Peatland Cover Changes Using Support Vector Machine and Markov Chain Model. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *14*(1), 294-301. https://doi.org/10.12928/telkomnika.v14i1.2400

Khoiriyah, Y. M., & Sitanggang, I. S. (2014, October). A spatial decision tree based on topological relationships for classifying hotspot occurences in Bengkalis Riau Indonesia. In *Advanced Computer Science and Information Systems (ICACSIS), 2014 International Conference on* (pp. 268-272). IEEE. https://doi.org/10.1109/ICACSIS.2014.7065844

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Nurpratami, I. D., & Sitanggang, I. S. (2015). Classification rules for hotspot occurrences using spatial entropy-based decision tree algorithm. *Procedia Environmental Sciences*, *24*, 120-126. https://doi.org/10.1016/j.proenv.2015.03.016

Nurulhaq, N. Z., & Sitanggang, I. S. (2015, August). Sequential Pattern Mining on hotspot data in Riau province using the PrefixSpan algorithm. In *Adaptive and Intelligent Agroindustry (ICAIA), 2015 3rd International Conference on* (pp. 257-260). IEEE. https://doi.org/10.1109/ICAIA.2015.7506517

Siknun, G. P., & Sitanggang, I. S. (2016). Web-based classification application for forest fire data using the shiny framework and the C5. 0 algorithm. *Procedia Environmental Sciences*, *33*, 332-339. https://doi.org/10.1016/j.proenv.2016.03.084

Sitanggang, I. S., & Fatayati, E. (2016). Mining Sequence Pattern on Hotspot Data to Identify Fire Spot in Peatland. *International Journal of Computing & Information Sciences*, *12*(1), 143. https://doi.org/10.21700/ijcis.2016.117

Sitanggang, I. S., Yaakob, R., Mustapha, N., & Ainuddin, A. N. (2014). A decision tree based on spatial relationships for predicting hotspots in peatlands. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *12*(2), 511-518.

https://doi.org/10.12928/telkomnika.v12i2.68 https://doi.org/10.12928/TELKOMNIKA.v12i2.2036

Sitanggang, I. S., Yaakob, R., & Mustapha, N. (2015). Burn area processing to generate false alarm data for hotspot prediction models. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *13*(3), 1037-1046. https://doi.org/10.12928/telkomnika.v13i3.1543

Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. Boston. US: Pearson Addison Wesley; 2005.

Thariqa, P., Sitanggang, I. S., & Syaufina, L. (2016). Comparative Analysis of Spatial Decision Tree Algorithms for Burned Area of Peatland in Rokan Hilir Riau. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *14*(2), 684-691. https://doi.org/10.12928/telkomnika.v14i2.3540

Pandya, R., & Pandya, J. (2015). C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications*, *117*(16). https://doi.org/10.5120/20639-3318

Primajaya, A., Sitanggang, I. S., & Syaufina, L. (2017, January). Visualization of spatial decision tree for predicting hotspot occurrence in land and forest in Rokan Hilir District Riau. In *IOP Conference Series: Earth and Environmental Science* (Vol. 54, No. 1, p. 012055). IOP Publishing. https://doi.org/10.1088/1755-1315/54/1/012055

Wijaya, P. T., Sitanggang, I. S., & Syaufina, L. (2016). Density Based Clustering of Hotspots in Peatland with Road and River as Physical Obstacles. *Indonesian Journal of Electrical Engineering and Computer Science*, *3*(3), 714-720. https://doi.org/10.11591/ijeecs.v3.i3.pp714-720