(5)

Proceedings
2010 Eighth International Conference on ICT and Knowledge Engineering

---

November 24-25, 2010 Bangkok, Thailand

IEEE Catalog Number: **CFP1028H-PRT**
ISBN: **978-1-4244-9875-8**
ISSN: **2157-0981**
IEEE Conference: **#17100**

Jointly organized by

SIAM UNIVERSITY

APD MEN

IEEE

Hotspot Occurrences Classification using Decision Tree Method

Case Study in the Rokan Hilir, Riau Province, Indonesia

Imas Sukaesih Sitanggang
Computer Science Department
Bogor Agricultural University
Bogor, Indonesia
e-mail: imas.sitanggang@ipb.ac.id

Mohd. Hasmadi Ismail
Forest Surveying & Engineering Laboratory
Faculty of Forestry
University Putra Malaysia
email: mhasmadi@putra.upm.edu.my

Abstract—**Application of geospatial and data mining techniques in forest fires research have resulted interesting and useful information in decision making related to the forest fires management. This paper presents a result of the study in applying the C4.5 algorithm on a forest fire dataset in the Rokan Hilir district, Riau Province, Indonesia. The dataset consists of hotspot occurrence locations, human activity factors, and land cover types. Human activity factors include city center locations, roads network and rivers network. The results were a decision tree which contains 18 leaves and 26 nodes with accuracy about 63.17%. Most of positive examples (the area with hotspot occurrences) and negative examples (no hotspot occurrences in the area) that are incorrectly classified by the model are located near rivers and roads.**

Keywords- **C4.5 algorithm, hotspot occurrences, decision tree method**

I. INTRODUCTION

Computer systems are capable to collect a huge spatial data that lead to a remarkable interest in applying data mining techniques. Some data mining tasks includes an association rules mining, classification and prediction, as well as cluster analysis have been successfully applied in analyzing spatial data related to many areas such as forest fires. A study by [1] used clustering and Hough transformation to reduce false alarm from the sets of hotspots in forest fire regions derived from NOAA imagery. Meanwhile [2] utilized classification algorithms including logistic regression and decision trees (J48), random forests, bagging and boosting of decision trees to develop predictive models of hotspot occurrences based on the forest structure using a GIS (geographical information system), meteorological ALADIN data and MODIS satellite data. Reference [3] proposed the incremental association mining method to obtain the primitive estimation of the fire grade from the historical fire data. The clustering algorithm K-means together with fuzzy logic has been applied by [4] to determine the fire risk spots from spatial data. The association rule algorithm namely Apriori was applied to analyze the probability and intensity of the forest fire effectively with coarse forest fire data in the forest area in the south of Beijing, China [5].

This study was carried out to extract a forest fire data and classifying hotspots occurrences by utilizing the decision tree algorithm namely C4.5. We developed a classification model for hotspot occurrences based on human activity factors including location of city center, road network, river network and land cover types in the Rokan Hilir district, Riau Province, Indonesia. This paper is structured as follows; Section 1 is an introduction, Section 2 describes the decision tree method and the C4.5 algorithm. Spatial data used in this study are briefly explained in Section 3. In section 4 we present and discuss the experiment results. Error visualization of the model is discussed in Section 5. Finally, we summarize the paper in Section 6.

II. DECISION TREE AND C4.5 ALGORITHM

A decision tree is a tree structure, in which each internal node (nonleaf node) denotes a test condition on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The root node, the topmost node, and internal node contain attribute test conditions to separate tuples into some partitions. A rule obtained from a decision tree consists of test attributes and their value in tree paths starting from the root node to the leaves node (terminals). Information Gain is generally used to determine the splitting attribute for the root node and internal nodes in a decision tree.

The most common algorithms for developing decision trees are Quinlan's ID3, C4.5 as a successor of ID3 and CART (Classification and Regression Tree). The ID3 algorithm compute the information gain for each attribute and select one that has the highest value. C4.5 is a successor of ID3 that learns decision tree classifiers. The following C4.5 algorithm generates a decision tree from a set D of cases [6]: (1) If D satisfies a stopping criterion, the tree for D is a leaf associated with the most frequent class in D. One reason for stopping is that D contains only cases of this class. (2) Some test T with mutually exclusive outcomes $T_1, T_2,\ldots,T_k$ is used to partition D into subsets $D_1, D_2, \ldots,D_k$, where D_i contains those cases that have outcome T_i. The tree for D has test T as its root with one subtree for each outcome T_i that is constructed by applying the same procedure recursively to the cases in D_i.

978-1-4244-9875-8/10/$26.00 ©2010 IEEE

The C4.5 algorithm visits each decision node recursively and selects optimal splitting attributes until the data set satisfies a stopping criterion. The recursion stops when either there is only one class remaining in the data, or there are no features left [7]. This algorithm uses Information Gain to select optimal splitting attributes, e.g., let a node N represents the tuples of partition D. The attributes with the highest information gain is chosen as the splitting attribute for the node N. This attribute minimizes the information needed to classify the tuples in resulting partitions and reflects the least randomness or "impurity" in these partitions [8]. The expected information needed to classify a tuple in D is given by

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$. Info(D) is the average amount of information needed to identify the class label of a tuple in D [8]. Info(D) is also known as the entropy of D. Assume that we want to partition the tuples in D on an attribute A having v distinct values, $\{a_1, a_2, \ldots, a_v\}$. The resulted partitions are related to the branches of the node N. $Info_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A [8].

$$Info_A(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} \times Info(D_j) \quad (2)$$

The term $\frac{|D_j|}{|D|}$ acts as the weight of the jth partition. Information gain is defined as the difference between the original information requirement (i.e. based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A) [8].

$$Gain(A) = Info(D) - Info_A(D) \quad (3)$$

The attribute A with the highest information gain, Gain (A), is chosen as the splitting attribute at node N.

In order to evaluate the performance of the classification model, a confusion matrix is calculated. The entries of matrix store number of test tuples predicted correctly and incorrectly by the model. The model accuracy is commonly used to determine the performance of the model. It is defined as [9]:

$$Accuracy = \frac{Number\ of\ correct}{Total\ number\ of\ p} \quad (4)$$

III. Dataset

The study area is located at the Rokan Hilir district in the Riau Province, Sumatera, Indonesia. The total area of Rokan Hilir is 896,142.93 ha. or approximately 10% of the total area of the Riau Province (8,915,015.09 ha.). It is situated in area between 100° 17' - 101° 21' East Longitude and 1° 14' - 2° 45' North Latitude. The spatial data of forest fire were preprocessed to prepare a dataset for mining purpose. There are two main tasks in constructing a forest fire dataset; firstly creating targets attribute and populating its value from the target objects (location of hotspots), and secondly creating explanatory attributes from neighbor objects related to the target objects. These steps were performed using open source tools: Quantum GIS 1.0.2 for spatial data analysis and visualization, PostgreSQL 8.4 as the spatial database management system (DBMS), and PostGIS 1.4 for spatial data analysis.

The target attribute contains positive and negative examples of hotspot occurrences. Positive examples are locations of hotspots along the year 2008 recorded by NOAA-18. Data containing locations of hotspots were obtained from the Ministry of Forestry, Republic of Indonesia. Negative examples are randomly generated and they are located within the area at least 1 km away from any positive examples. For this purpose we created 1 km buffer from positive examples and extracted by random a generated points outside the buffer to be negative examples. The forest fires dataset was analyzed using the J48 module as Java implementation of C4.5 in the data mining toolkit Weka 3.6.2.

IV. Results and Discussion

The dataset contains 744 tuples (374 positive examples and 370 negative examples). There were one target attribute (class of examples) and four explanatory attributes: (1) *min_dist_to_road*, (2) *min_dist_to_river*, (3) *min_dist_to_city* represent distance from the location of examples to nearest road, river and city center, (4) *land cover types* for area where the examples were located, respectively. The datasets were divided into two groups; training data to develop a classification model and testing data to calculate accuracy of the model. We applied the 10-*folds cross validation* [10] to determine accuracy of the classifier. The decision tree contains 18 leaves and 26 nodes with the first test attribute were land cover types. Below were some rules extracted from the tree:

1. IF landcovertype = Plantation AND min_dist_to_river <= 4546.97 meters THEN Hotspot Occurrence = F (187.0/76.0)
2. IF landcovertype = Plantation AND min_dist_to_river > 4546.97 meters THEN Hotspot Occurrence = T (125.0/30.0)
3. IF landcovertype = Swamp AND min_dist_to_road <= 3366.85 meters THEN Hotspot Occurrence = F (3.0)
4. IF landcovertype = Shrubs THEN Hotspot Occurrence = F (63.0/28.0)
5. IF landcovertype = Unirrigated_agricultural_field AND min_dist_to_river > 353.66 meters THEN Hotspot Occurrence = T (40.0/19.0)
6. IF landcovertype = Dryland_forest AND min_dist_to_city <= 14807.65 meters THEN Hotspot Occurrence = F (77.0/27.0)
7. IF landcovertype = Mix_garden AND min_dist_to_city <= 16354.78 meters THEN Hotspot Occurrence = F (65.0/15.0)

978-1-4244-9875-8/10/$26.00 ©2010 IEEE

8. IF landcovertype = Mix_garden AND min_dist_to_city > 16354.78 meters AND min_dist_to_city <= 23910.15 meters THEN Hotspot Occurrence = T (54.0/10.0)

The numbers (parentheses) at the end of each leaf represented the number of examples in this leaf whereas the number of misclassified examples were given after a / slash /. There were 470 (63.172 %) instances (tuples) that were correctly classified by the tree. The classification model can be used to predict the hotspot occurrences at the new location. To show how this task was performed we generate randomly a total of 165 points that were not exist in the dataset. Figure 1 showed that a point 187 was located in plantation area with the distance to nearest river was 6.09 km. According to Rule 2 having the body "landcovertype = Plantation AND min_dist_to_river > 4,546.97 meters", this point is classified into a positive example (fire occurrences is True).

The accuracy of classification is still low (63.172 %). This due to 9 rules of 18 rules is supported by small fraction of the data (below 2.5% of the overall records). As an example there are only three records that support Rule 3 (IF landcovertype = Swamp AND min_dist_to_road <= 3366.85 meters THEN Hotspot Occurrence = F). This situation may result incorrect classes when the decision tree is applied to the test set and then may decrease the accuracy. In order to improve accuracy of the classifier, the following data preprocessing tasks can be performed: 1) identify outliers and smooth out noisy data; 2) attribute transformation and discretization; and 3) attribute selection.

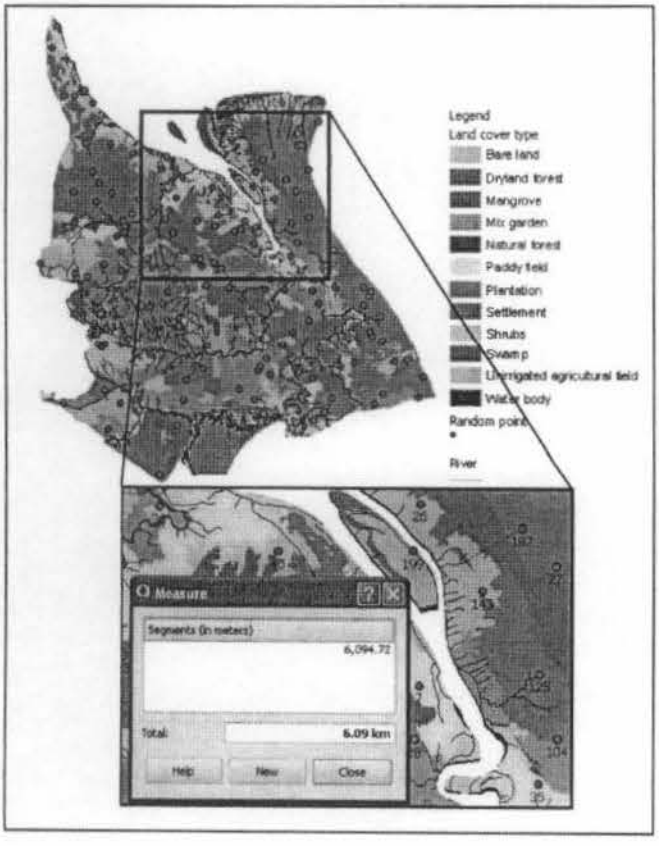


Figure 1. Distance from location (point 187) to a nearest river.

V. Error Visualization

Figure 2 visualized classification error of the decision tree in which the axis Actual class represents class labels in the dataset and Predicted class indicates class labels predicted by the model. Records with class label T (True) were positive examples (the area with hotspot occurrences), whereas records with class label F (False) were negative examples (No hotspot occurrences in the area). The correctly classified records were indicated by crosses and incorrectly classified ones were showed up as squares. There were 171 positive examples or 22.98% with class label True were predicted as negative examples (class label False). These records were indicated as blue squares in Figure 2. Red squares in Figure 2 depicted 103 negative examples (13.84%) with class label False that were predicted as positive examples (class label True). Figure 3 indicated that number of correctly and incorrectly classified records grouped by land cover types. The study area is largely cover by Plantation (312 records or 41.94 % of the overall records) in which 205 records were correctly classified and 107 records were incorrectly classified by the model.

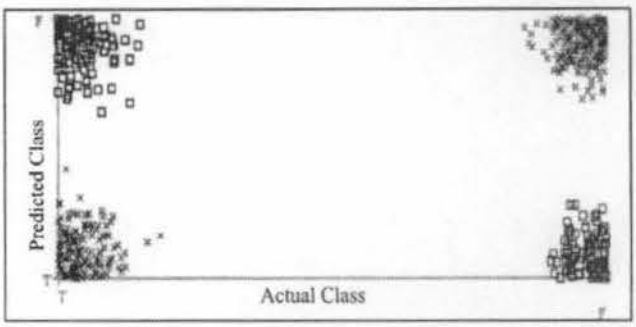


Figure 2. Error visualization.

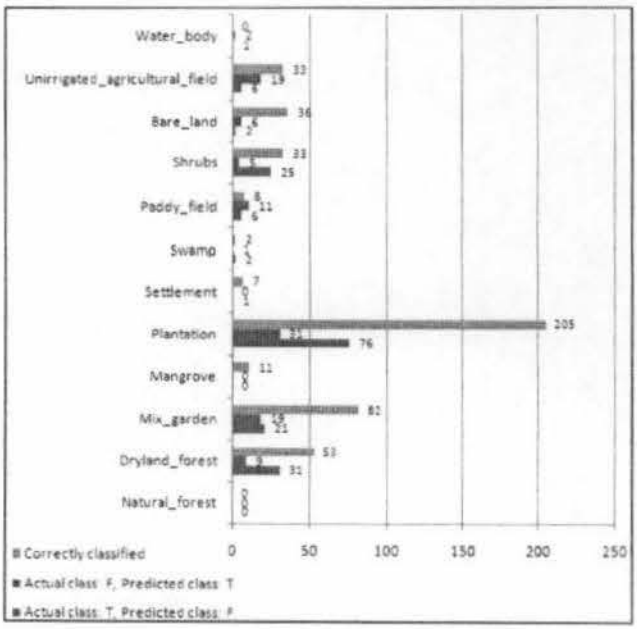
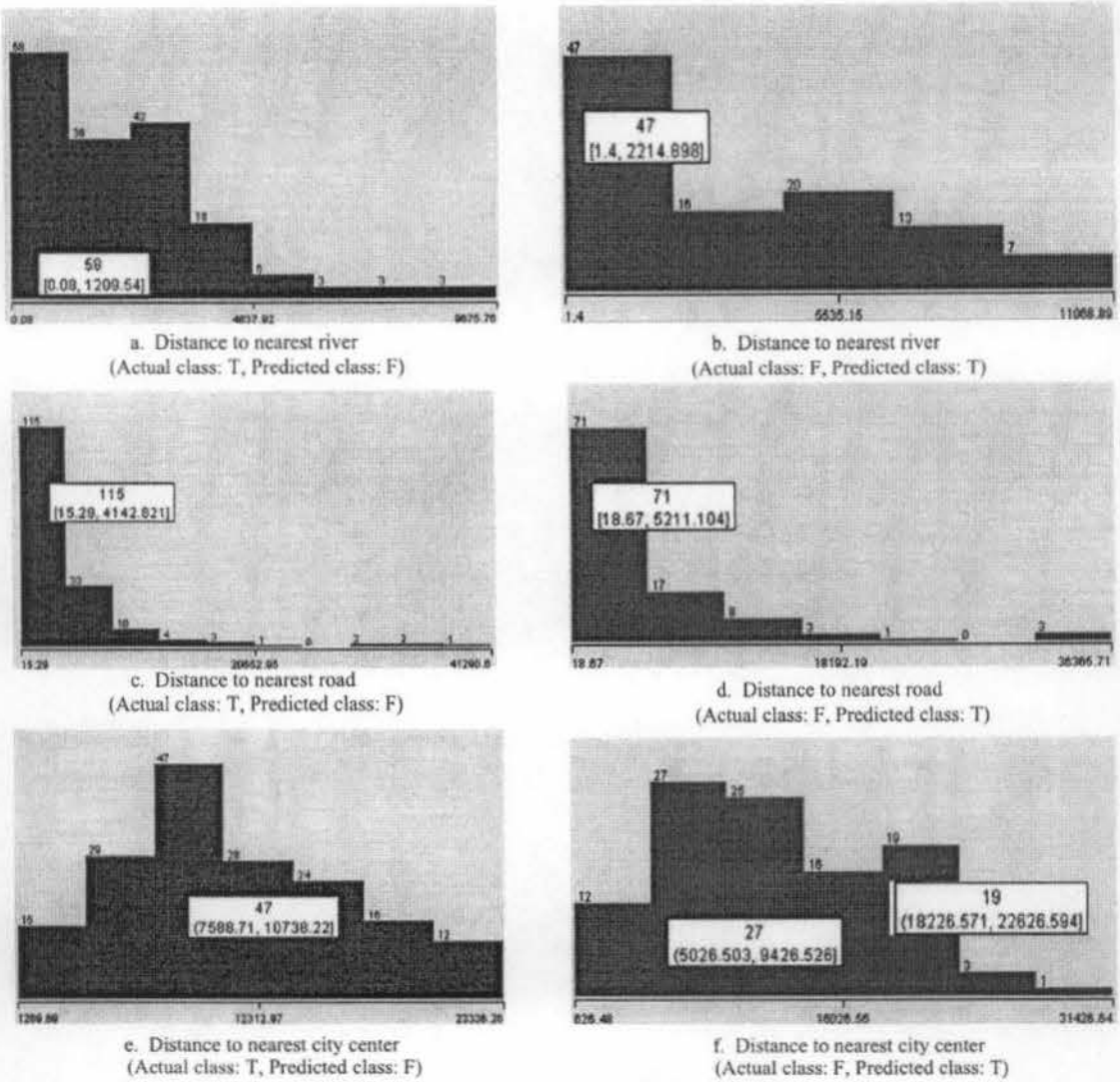


Figure 3. Number of correctly and incorrectly classified records grouped by land cover types.

Figure 4 shows a number of incorrectly classified records grouped by distance to nearest river, road and city center. Most positive and negative examples that were incorrectly classified by the model were located near the river and roads (Figure 4a, 4b, 4c, and 4d). From Figure 4a, it can be stated that a total of 58 positive examples (class label True) with distance to nearest river in the interval [0.08 – 1209.54] meters are predicted as negative examples (class label False).



a. Distance to nearest river
(Actual class: T, Predicted class: F)

b. Distance to nearest river
(Actual class: F, Predicted class: T)

c. Distance to nearest road
(Actual class: T, Predicted class: F)

d. Distance to nearest road
(Actual class: F, Predicted class: T)

e. Distance to nearest city center
(Actual class: T, Predicted class: F)

f. Distance to nearest city center
(Actual class: F, Predicted class: T)

Figure 4. Histogram for incorrectly classified records.

Despite records indicated the actual class True but classified as the class False spread throughout the study area near or far from city centers (Figure 4e). There are 47 records are located in the area where the distance to nearest city centers is in the interval from 7588.71 – 10738.22 meters. While Figure 4f shows that majority of records (96.12%) are located in area with distance to nearest city center less than or equal to 22626.594 meters. These records have the actual class False and they are incorrectly classified by the model as the class True.

VI. SUMMARY

This study applied the C4.5 algorithm in developing a decision tree to classify forest hotspot occurrences in the Rokan Hilir District, Riau Province Indonesia. The target attribute contains positive examples (hotspot occurrences) and negative examples whereas explanatory attributes related to human activity factors i.e. the location of city center, road network, river network and land cover types. There were 18 classification rules generated from the tree with the accuracy of 63.17 %. The result showed that there were 171 positive examples (22.98%) with class label True (the area with hotspot occurrences) was predicted as

negative examples with class label False (No hotspot occurrences in the area). Most of positive and negative examples that are incorrectly classified by the model are located near rivers and roads. The decision tree algorithm namely C4.5 and data mining method used in this study is useful as predictive model for forest fire occurrences.

ACKNOWLEDGMENT

This work is supported by the Indonesia Directorate General of Higher Education (IDGHE), Ministry of National Education, Indonesia.

REFERENCES

[1] S. C. Tay, W. Hsu, K. H. Lim, and L. C. Yap, "Spatial data mining: clustering of hot spots and pattern recognition," IEEE, pp. 3685-3687, 2003.

[2] D. Stojanova, P. Panov, A. Kobler, S. Džeroski, and K. Taškova, "Learning to predict forest fires with different data mining techniques," Conference on Data Mining and Data Warehouses (SiKDD 2006), Ljubljana, Slovenia, 2006.

[3] L. Yu and F. Bian, "An incremental data mining method for spatial association rule in GIS based fireproof system," IEEE, 2007.

[4] K. S. N. Prasad and S. Ramakrishna, "An autonomous forest fire detection system based on spatial data mining and fuzzy logic," IJCSNS International Journal of Computer Science and Network Security, vol.8 no.12, pp. 49-55, December 2008.

[5] H. Lin, Z. Goumin, and Q. Yun, "Application of apriori algorithm to the data mining of the wildfire," Sixth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE, pp. 426-429, 2009.

[6] J. R. Quinlan, "Improved Use of Continuous Attributes in C4.5," Journal of Artificial Intelligence Research 4, pp. 77-90, 1996.

[7] S. Marsland, Machine Learning, An Algorithmic Perspective, CRC Press, Taylor & Francis Group, 2009.

[8] J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan-Kaufmann: San Diego, USA, 2006.

[9] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Pearson Education Inc, 2006.

[10] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Series: San Francisco, 2005.

978-1-4244-9875-8/10/$26.00 ©2010 IEEE