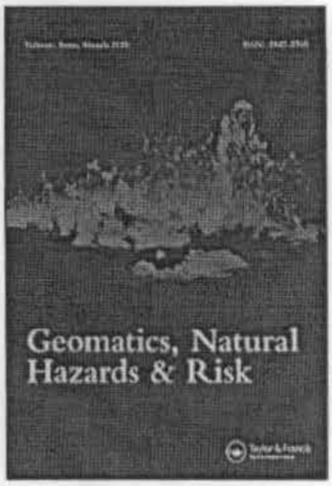
14 12



Geomatics, Natural Hazards and Risk
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t913444127

Classification model for hotspot occurrences using a decision tree method

Imas Sukaesih Sitanggang[a]; Mohd Hasmadi Ismail[b]
[a] Department of Computer Science, Bogor, Indonesia [b] Forest Surveying and Engineering Laboratory, Faculty of Forestry, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

Online publication date: 19 May 2011

To cite this Article Sitanggang, Imas Sukaesih and Ismail, Mohd Hasmadi(2011) 'Classification model for hotspot occurrences using a decision tree method', Geomatics, Natural Hazards and Risk, 2: 2, 111 — 121
To link to this Article: DOI: 10.1080/19475705.2011.565807
URL: http://dx.doi.org/10.1080/19475705.2011.565807

Geomatics, Natural Hazards and Risk
Vol. 2, No. 2, June 2011, 111–121

Taylor & Francis
Taylor & Francis Group

Classification model for hotspot occurrences using a decision tree method

IMAS SUKAESIH SITANGGANG† and MOHD HASMADI ISMAIL*‡

†Department of Computer Science, Bogor Agricultural University, Bogor 16680, Indonesia

‡Forest Surveying and Engineering Laboratory, Faculty of Forestry, Universiti Putra Malaysia, Serdang 43400 UPM, Selangor, Malaysia

(*Received 11 November 2010; in final form 20 February 2011*)

Forest fires in Indonesia mostly occur because of errors or bad intentions. This work demonstrates the application of a decision tree algorithm, namely the C4.5 algorithm, to develop a classification model from forest fire data in the Rokan Hilir district, Indonesia. The classification model used is a collection of IF-THEN rules that can be used to predict hotspot occurrences for forest fires. The spatial data consist of the location of hotspot occurrences and human activity factors including the location of city centres, road and river networks as well as land cover types. The results were a decision tree containing 18 leaves and 26 nodes with an accuracy of 63.17%. Each leaf node holds positive and negative examples of hotspot occurrences whereas the root and internal nodes contain attribute test conditions: the distance from the location of examples to the nearest road, river, city centre and the land cover types for the area where the examples are located. Positive examples are hotspot locations in the study area and negative are randomly generated points within the area at least 1 km away from any positive example. The classification model categorized whether the region was susceptible to hotspots occurrences or not. The model can be used to predict hotspot occurrences in new locations for fire prediction.

1. Introduction

Forest fires in Indonesia are not a new phenomenon and seem to be a yearly tradition especially in the dry season. Almost 100% of forest and land fires in Indonesia are caused by human-made factors both on purpose and accidently (Lailan 2008), whereas natural factors such as lightning, volcano eruption and burned coal have insignificant influences on the fires compared to human factors. Fires are considered as a cheap way to clear land for new plantations and also to extract natural resources, such as harvesting timber in forests. Forest fires lead to several environmental problems including loss of forest as carbon sink, loss of biodiversity, transboundary haze pollution, as well as significant economic losses (Danan 2008). Fire prevention is important for solving forest fire problems in Indonesia in order to avoid significant material damages to the natural environment. Therefore study related to the identification of forest fire risk is necessary. Many works have been conducted to develop forest fire risk models for regions in Indonesia (Jaruntorn 2001, Mulyanto

*Corresponding author. Email: imas.sitanggang@gmail.com *et al.* 2001, Mustara 2006, Danan 2008). Geographic information systems (GISs) and remote sensing have been used to analyse forest fire data (Mulyanto *et al.* 2001, Mustara 2006, Danan 2008). In addition to GISs and remote sensing, the Complete Mapping Analysis method (CMA) (Jaruntorn 2001, Mustara 2006) and Multi-criteria Analysis (MCA) (Danan 2008) are also applied to understand the causes of fire risk factors and interactions between them.

Huge amounts of spatial data in many areas have been collected in various computer systems. Nowadays spatial databases, as one of the components of GISs, store large numbers of spatial features and their relationships for further manipulation and analysis help users in decision-making processes. This situation has led to an increase in applying data mining techniques to extract the interesting and useful but implicit spatial patterns from large numbers of spatial data. Some functionality in data mining, including association rules mining, classification and prediction, as well as cluster analysis, have been successfully applied for analysing spatial data related to forest fires. The association rule algorithm namely *Apriori* was applied to analyse the probability and intensity of the forest fire effectively with coarse forest fire data in the forest area in the south of Beijing (Hu *et al.* 2009). Liang and Fuling (2007) 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, have been applied to determine spots at the risk of forest fire from spatial data (Kalli and Ramakrishna 2008). Another study by Seng *et al.* (2003) used clustering and Hough transformation to reduce false alarms from the set of hotspots in forest fire regions derived from National Oceanic and Atmospheric Administration (NOAA) images. Daniela *et al.* (2006) 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 GIS, meteorological ALADIN data and Moderate Resolution Imaging Spectroradiometer (MODIS) satellite data.

This paper presents the work of extracting beneficial information from forest fire data by utilizing a decision tree algorithm, namely C4.5. We develop a classification model for hotspot occurrences based on human activity factors including the location of city centres, road networks, river networks and land cover types in the Rokan Hilir district, Riau Province, Indonesia. The model provides descriptions for areas where hotspots occur. This information is important in planning fire fighting strategies. We can also predict hotspot occurrences in new areas given their human activity factors. Predicting hotspot occurrences is essential as an early warning system for preventing large forest fires and thus major damages can be avoided.

2. Spatial data

The study area is the Rokan Hilir district in the Riau Province, Sumatra Indonesia (figure 1). 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.

There are two categories of the spatial data: (1) a target object – location of hotspot occurred in 2008; and (2) neighbouring objects of the target object – location of the city centre, road network, river network and land cover type. All objects are represented in vector format. We assign the spatial reference system UTM 47N and datum WGS84 to all objects.

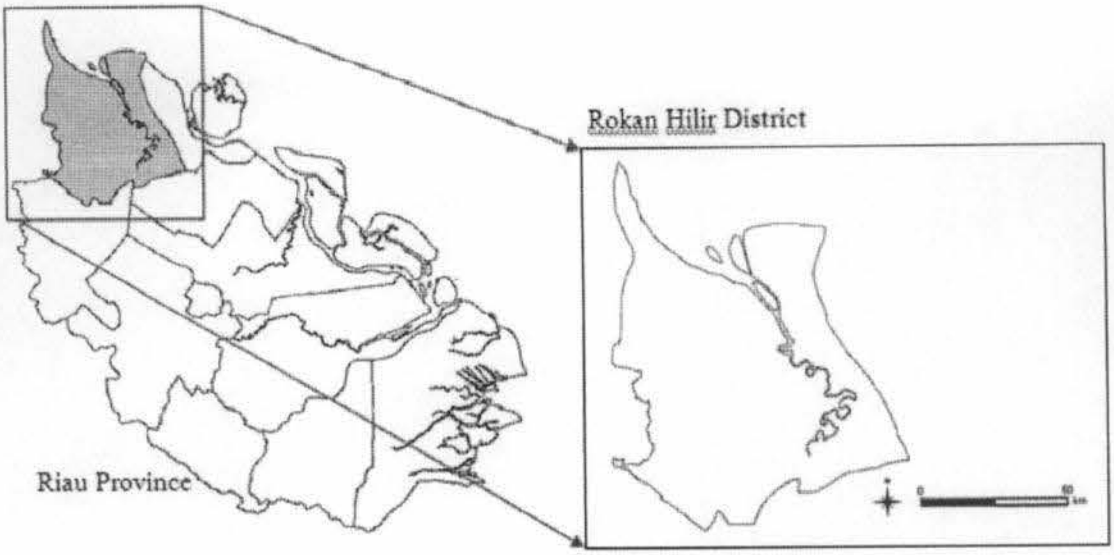


Figure 1. Rokan Hilir district in Sumatra, Indonesia.

3. Constructing a forest fire dataset

The spatial data are preprocessed to construct a dataset for the C4.5 algorithm. For mining purposes using the classification algorithms, a dataset should contain some explanatory attributes and one target attribute. There are two main tasks in constructing a forest fire dataset: (1) creating a target attribute and populating its value from the target object (location of hotspots); and (2) creating explanatory attributes from neighbouring objects related to the target object. 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, 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 throughout the year 2008 recorded by NOAA-18. Data containing the 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 create a 1 km buffer from positive examples and extract all randomly generated points outside the buffer to be negative examples. Figure 2 shows the distribution of positive and negative examples.

In order to create explanatory attributes we applied spatial relationship operators to relate neighbouring objects and the target object consisting of positive and negative examples of hotspot occurrences. The target object and neighbouring objects (river, road and city centre) are shown in figure 3. The distance between a location of example and the nearest river is calculated using the spatial relationship operator ST Distance in PostGIS and then the result is a value of the attribute *min_dist_to_river* in the dataset. Distances from locations of example are also determined to the nearest road and nearest city centre. Afterwards the results are stored respectively in the attribute *min_dist_to_road* and *min_dist_to_city*.

Another attribute *land_cover_type* stores types of land cover for the area where the examples of hotspot occurrences are located (figure 4). The spatial relationship operator ST_Within in PostGIS was applied to determine values of attribute *land_cover_type*.

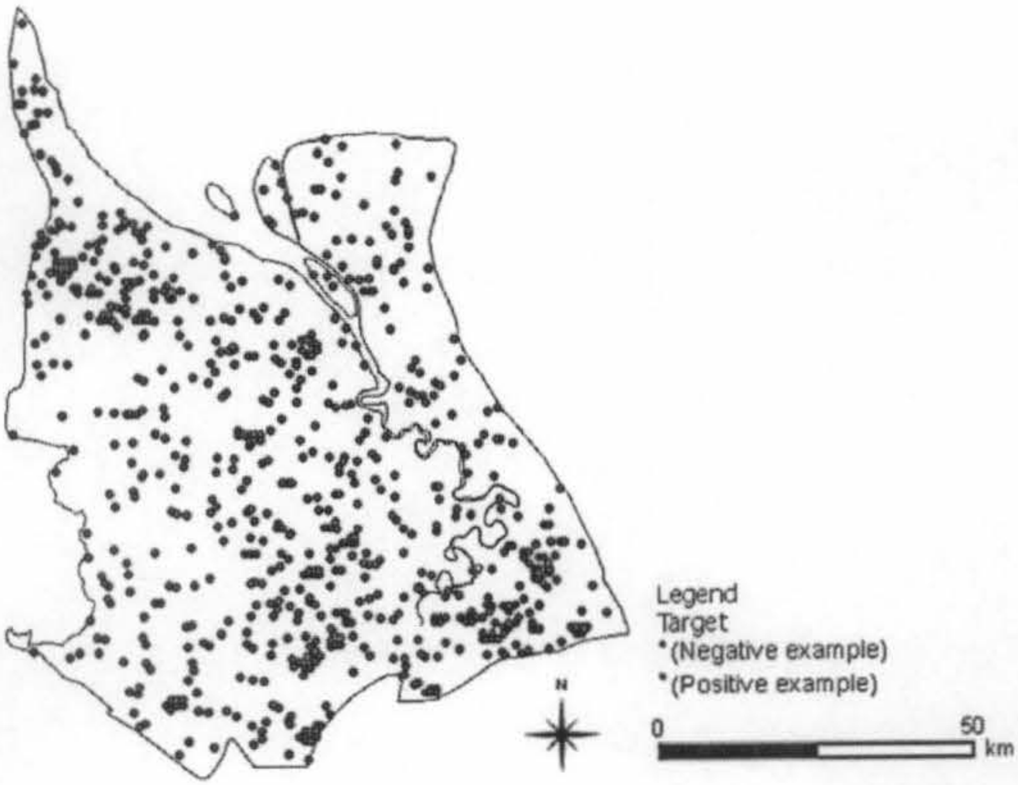


Figure 2. Positive and negative examples of hotspot occurrences.

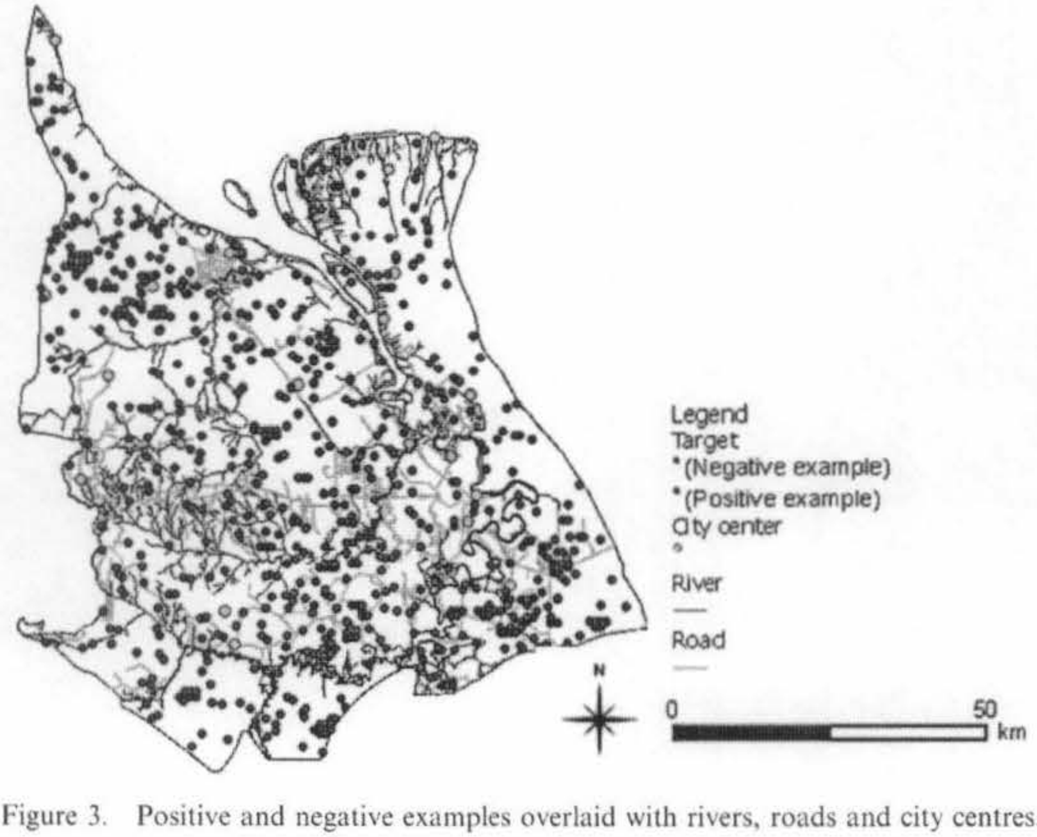


Figure 3. Positive and negative examples overlaid with rivers, roads and city centres.

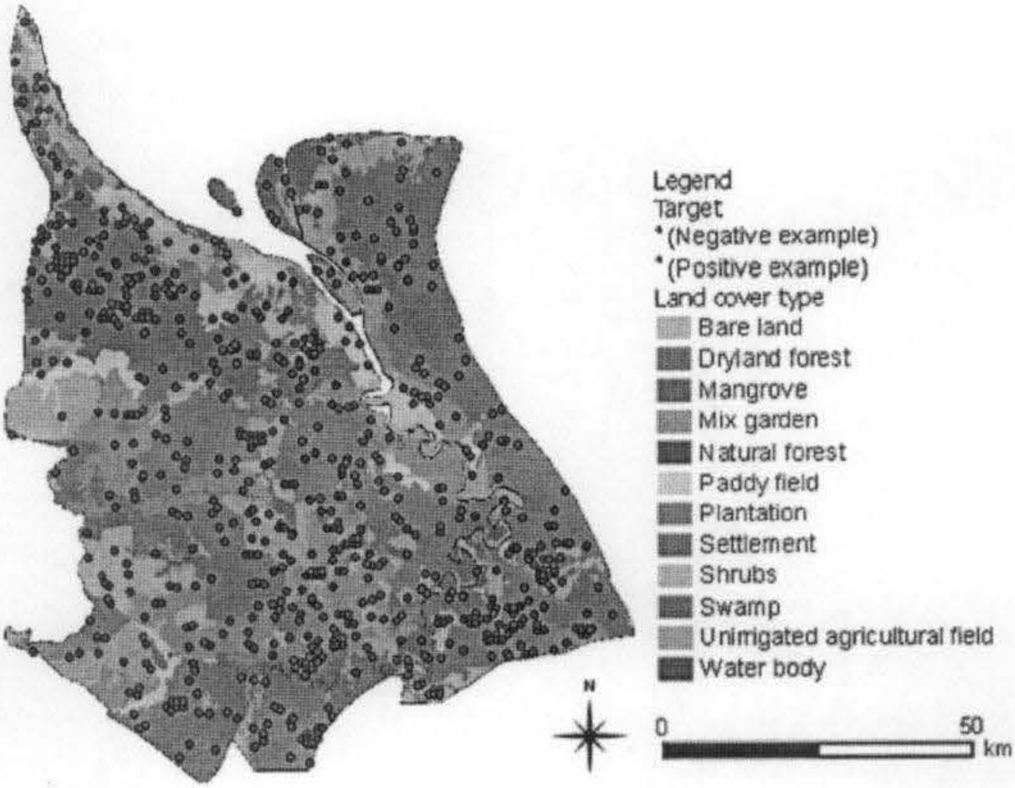


Figure 4. Positive and negative examples overlaid with land cover.

4. Decision tree algorithm

One of the data mining tasks is a classification that is widely used to extract models describing the data classes and predicting class labels for new data. In the classification task, we aim to discover classification rules that determine the class label of any object (Y) from the values of its attributes (X). Decision tree induction is a simple and powerful technique for extracting classification rules from class-labelled training tuples. 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 the 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.

Let a node N represent the tuples of partition D. The attribute 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 (Jiawei and Micheline 2006). The expected information needed to classify a tuple in D is given by

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

where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|\mathrm{C}_{i,\mathrm{D}}|/|\mathrm{D}|$. Info(D) is the average amount of information needed to identify the class label of a tuple in D (Jiawei and Micheline 2006). 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 resulting partitions are related to the branches of the node N. $\mathrm{Info}_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A (Jiawei and Micheline 2006).

$$\mathrm{Info}_{\mathrm{A}}(\mathrm{D}) = \sum\nolimits_{j=1}^{v} \frac{|D_j|}{|D|} \times \mathrm{Info}(\mathrm{D}_j) \tag{2}$$

The term $\frac{|D_j|}{|D|}$ acts as the weight of the *j*th partition. The 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) (Jiawei and Micheline 2006).

$$\mathrm{Gain}(\mathrm{A}) = \mathrm{Info}(\mathrm{D}) - \mathrm{Info}_{\mathrm{A}}(\mathrm{D}) \tag{3}$$

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

Decision tree algorithms are popular in knowledge discovery because they can handle high-dimensional data and the tree form of acquired knowledge is generally easily interpreted by humans. In addition, we can generate IF-THEN rules from the tree paths starting from the root node to the leaf node (terminals). A rule consists of test attributes in its body part and a target class as the head part of the rule. The collection of IF-THEN rules is relatively easy to implement in developing a classification system using a programming language. The most common algorithms for developing decision trees are Quinlan's ID3, the C4.5 as a successor of ID3 and CART (Classification and Regression Tree). The ID3 algorithm computes the information gain for each attribute and selects one that has the highest value. The C4.5 is a successor of ID3 that learns decision tree classifiers. This algorithm visits each decision node recursively and selects optimal splitting attributes until the dataset satisfies a stopping criterion. The recursion stops when either there is only one class remaining in the data, or there are no features left (Stephen 2009). The C4.5 algorithm also uses Information Gain to select optimal splitting attributes. This algorithm uses a different method called rule post-pruning. There are three main tasks in C4.5: (1) generate the tree using the ID3 algorithm, (2) convert the tree to a set of IF-THEN rules, and (3) prune each rule by removing preconditions if the accuracy of the rule increases without it (Stephen 2009). Nowadays the C4.5 has probably become the most commonly used and studied decision tree algorithm. The results of Tjen-Sien *et al.* (2000) show that among decision tree algorithms, the C4.5 provides good combinations of error rate and speed.

Classification and Regression Trees is another commonly used algorithm for the induction of decision trees for classification proposed by Brieman *et al.* (1984). The basic methodology of divide and conquer, described in the C4.5, is also used in the CART. The differences are in the tree structure, the splitting criteria, the pruning method, and the way missing values are handled (Kohavi and Quinlan 1999). The CART constructs binary decision trees and branches on a single attribute-value pair rather than on all values of the selected attribute. In a particular case, the binary tree may be less interpretable with multiple splits occurring on the same attribute at adjacent levels. In developing a tree using CART, there may be no good binary split on an attribute that has a good multi-way split (Kononenko 1995). In our study, attributes in a spatial dataset have many district values that will become outcomes of the test attributes. For example land cover type is classified into some categories including plantation, swamp, shrubs, bare land, dryland forest and so on. These categories merged with spatial relations form two branches from the node. The left branch contains objects related to one of these categories, for example plantation, while objects in the right branch are associated to all other categories. Characteristics of objects in the right branch may not be specific because this node is related to some different categories of land cover rather than a single category. In this case the right branch may be extended to the next levels of the tree. This process may construct a less interpretable spatial decision tree because multiple splits on the same test attributes may occur at more than one level. For that we select the C4.5 algorithm that provides multi-way splits instead of the CART algorithm that applies binary splits in constructing the tree.

In order to evaluate the performance of the classification model, a confusion matrix is calculated (Pang-Ning *et al.* 2006). The entries of the matrix store a 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 (Pang-Ning *et al.* 2006):

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{4}$$

5. Results and discussion

The forest fires dataset was analysed using the J48 module as a Java implementation of C4.5 in the data mining toolkit Weka 3.6.2. The J48 package is a Weka's implementation of the decision tree learner. The package is a directory containing a collection of related classes that builds a C4.5 decision tree. Some classes included in the package are classes for computing the entropy and the information gain, and a class implementing a binary C4.5-like split on an attribute.

The dataset contains 744 tuples (374 positive examples and 370 negative examples). There is 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* representing the distance from the location of examples to the nearest road, river and city centre, respectively, (4) *land cover type* for the area where the examples are located. The dataset was divided into two groups: training data to develop a classification model and testing data to calculate the accuracy of the model. We applied the 10-folds cross validation (Ian and Eibe 2005) to determine the accuracy of the classifier. The decision tree contains 18 leaves and 26 nodes where the first test attribute is land cover type. Below are the rules extracted from the tree:

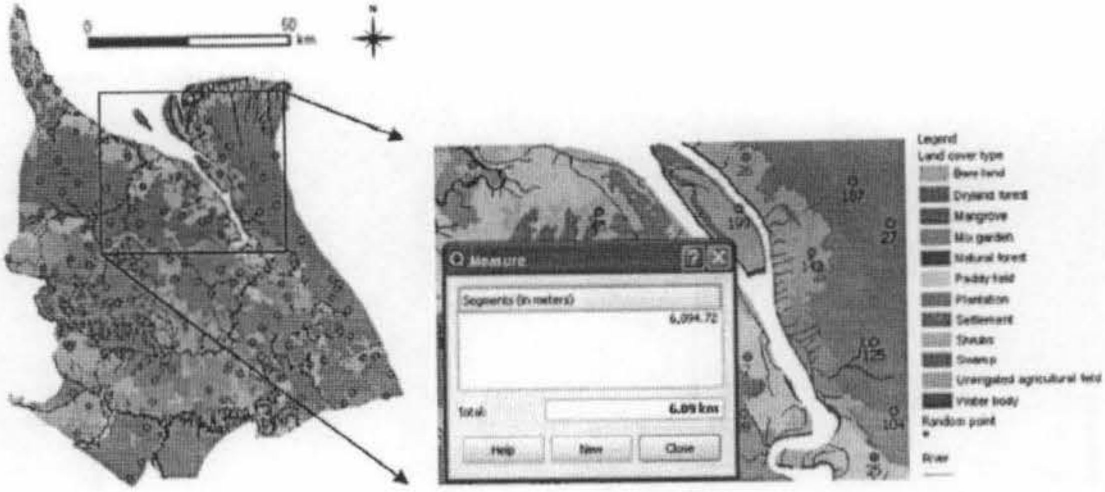
1. IF landcovertype = Plantation AND min_dist_to_river < = 4546.97 metres THEN Hotspot Occurrence = F (187.0/76.0)
2. IF landcovertype = Plantation AND min_dist_to_river > 4546.97 metres THEN Hotspot Occurrence = T (125.0/30.0)
3. IF landcovertype = Swamp AND min_dist_to_road < = 3366.85 metres THEN Hotspot Occurrence = F (3.0)

4. IF landcovertype = Swamp AND min_dist_to_road > 3366.85 metres THEN Hotspot Occurrence = T (2.0)
5. IF landcovertype = Shrubs THEN Hotspot Occurrence = F (63.0/28.0)
6. IF landcovertype = Bare_land THEN Hotspot Occurrence = T (44.0/8.0)
7. IF landcovertype = Unirrigated_agricultural_field AND min_dist_to_river < = 353.66 metres AND min_dist_to_road < = 77.59 metres THEN Hotspot Occurrence = T (2.0)
8. IF landcovertype = Unirrigated_agricultural_field AND min_dist_to_river < = 353.66 metres AND min_dist_to_road > 77.59 metres THEN Hotspot Occurrence = F (16.0)
9. IF landcovertype = Unirrigated_agricultural_field AND min_dist_to_river > 353.66 metres THEN Hotspot Occurrence = T (40.0/19.0)
10. IF landcovertype = Dryland_forest AND min_dist_to_city < = 14807.65 metres THEN Hotspot Occurrence = F (77.0/27.0)
11. IF landcovertype = Dryland_forest AND min_dist_to_city > 14807.65 metres THEN Hotspot Occurrence = T (16.0/3.0)
12. IF landcovertype = Settlement THEN Hotspot Occurrence = F (8.0/1.0)
13. IF landcovertype = Mangrove THEN Hotspot Occurrence = F (11.0)
14. IF landcovertype = Mix_garden AND min_dist_to_city < = 16354.78 metres THEN Hotspot Occurrence = F (65.0/15.0)
15. IF landcovertype = Mix_garden AND min_dist_to_city > 16354.78 metres AND min_dist_to_city < = 23910.15 metres THEN Hotspot Occurrence = T (54.0/10.0)
16. IF landcovertype = Mix_garden AND min_dist_to_city > 23910.15 metres THEN Hotspot Occurrence = F (3.0)
17. IF landcovertype = Paddy_field THEN Hotspot Occurrence = T (25.0/12.0)
18. IF landcovertype = Water Body THEN Hotspot Occurrence = F (3.0/1.0)

The numbers in (parentheses) at the end of each leaf represent the number of examples in this leaf whereas the number of misclassified examples are given after a / slash /. There are 470 (63.172%) instances (tuples) that are correctly classified by the tree. The classification model can be used to predict the hotspot occurrences on the new location. To show how this task can be performed we generate randomly 165 points that do not exist in the dataset. Figure 5 shows that a point 187 is located in the plantation where the distance to the nearest river is 6.09 km. According to Rule 2 having the body 'landcovertype = Plantation AND min_dist_to_river > 4,546.97 metres', this point is classified into a positive example (fire occurrence is true).

Figure 6 shows a point 64 located in swamp area where the distance to the nearest road is 3.03 km. According to Rule 3 having the body 'landcovertype = Swamp AND min_dist_to_road < = 3,366.85 metres', this point is classified into a false example (no fire occurrence). In addition the points 9, 67, 145 are located in shrubs thus from Rule 5 these points are classified into false examples (no hotspot occurrences).

Another example is given in figure 7 that shows point 128 located in mix garden where the distance to the nearest city centre is 19.97 km. According to Rule 15 with the body 'landcovertype = Mix_garden AND min_dist_to_city > 16,354.78 metres AND min_dist_to_city < = 23,910.15 metres', this point is classified into a true example (fire occurrence is true).



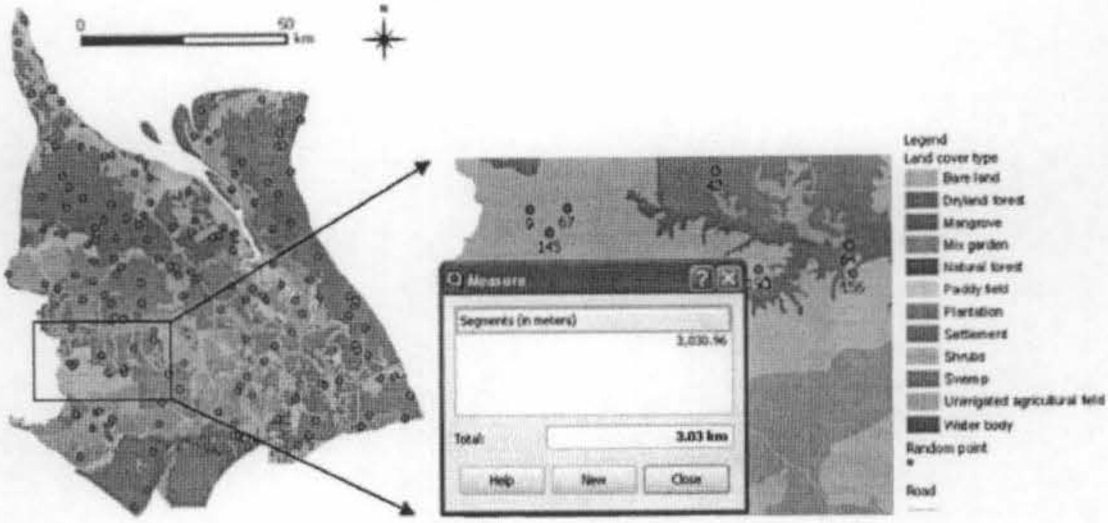
Measure

Segments (in meters)

6,094.72

Total: 6.09 km

Figure 5. Distance from a location (point 187) to the nearest river.



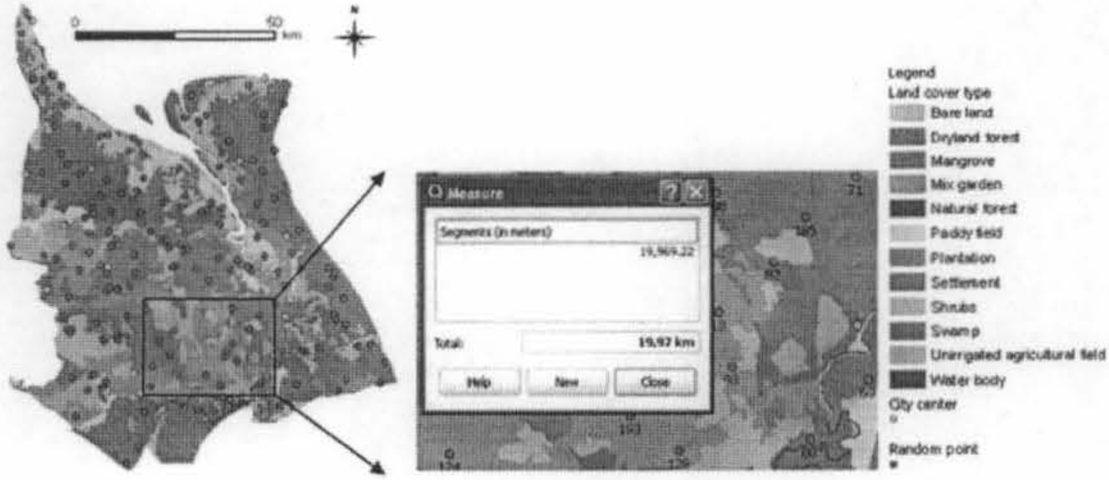
Measure

Segments (in meters)

3,030.96

Total: 3.03 km

Figure 6. Distance from a location (point 64) to the nearest road.



Measure

Segments (in meters)

19,969.22

Total: 19.97 km

Figure 7. Distance from a location (point 128) to the nearest city centre.

6. Conclusion

This work applied the C4.5 algorithm to develop 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 centres, road networks, river networks and land cover types. There are 18 classification rules generated from the tree with an accuracy of 63.17%. The use of the C4.5 algorithm on the forest fire dataset results in the decision tree predicting fire occurrences and describing characteristics of areas where the fires occur. The tree structure can easily be converted to a set of simple IF-THEN classification rules as a classification model for hotspot occurrences. The model validation should be performed by testing the model on new areas of forest fires. Predicting hotspot occurrences is important for wildfire prevention and determining fighting strategies. This work significantly shows the benefits of applying data mining techniques, namely a decision tree, on forest fire data to develop a classification model for predicting hotspot occurrences.

Acknowledgment

The authors are grateful to the anonymous reviewers for their comments/suggestions that have helped to improve an earlier version of the paper.

References

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., 1984, *Classification and Regression Trees* (Boca Raton, FL: Chapman & Hall/CRC Press).

Danan, P.H., 2008, A RS/GIS-based multi-criteria approaches to assess forest fire hazard in Indonesia (case study: West Kutai District, East Kalimantan Province). Master thesis, Bogor Agricultural University, Indonesia.

Daniela, S., Panče, P., Andrej, K., Sašo, D. and Katerina, T., 2006, Learning to predict forest fires with different data mining techniques. In *Conference on Data Mining and Data Warehouses* (SiKDD 2006), Ljubljana, Slovenia.

Hu, L., Zhou, G. and Qiu, Y., 2009, Application of a priori algorithm to the data mining of the wildfire. In Sixth International Conference on Fuzzy Systems and Knowledge Discovery, (New York: IEEE Press), pp. 426–429.

Ian, H.W. and Eibe, F., 2005, *Data Mining: Practical machine learning tools and techniques* (San Francisco, CA: Morgan Kaufmann Series).

Jaruntorn, B., 2001, GIS-based method in developing wildfire risk model (a case study in Sasamba, East Kalimantan, Indonesia). Master thesis, Bogor Agricultural University, Indonesia.

Jiawei, H. and Micheline, K., 2006, *Data Mining Concepts and Techniques* (San Diego, CA: Morgan-Kaufmann).

Kalli, S.N.P. and Ramakrishna, S., 2008, An autonomous forest fire detection system based on spatial data mining and fuzzy logic. *IJCSNS International Journal of Computer Science and Network Security*, 8, pp. 49–55.

Kohavi, R. and Quinlan, R., 1999, *Decision Tree Discovery*. Available online at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.5353 (accessed 30 August 2010).

Kononenko, I., 1995, A counter example to stronger version of the binary tree hypothesis. In *'ECML-95 Workshop on Statistics, Machine Learning and Knowledge Discovery in Databases*, Heraklion, Greece, April, pp. 31–36.

LAILAN, S., 2008, Kebakaran Hutan dan Lahan di Indonesia: Perilaku Api, Penyebab, dan Dampak Kebakaran (Malang: Bayumedia) (in Bahasa).

LIANG, Y. and FULING, B., 2007, An incremental data mining method for spatial association rule in GIS based fireproof system. In International Conference on Wireless Communications, Networking and Mobile Computing, 21–25 September 2007, Shanghai (New York: IEEE), pp. 5983–5986.

MULYANTO, D., MASAMU, A. and SATOSHI, T., 2001, Forest fire hazard model using remote sensing and geographic information systems: toward understanding of land and forest degradation on lowland areas of East Kalimantan Indonesia. In *22nd Asian Conference on Remote Sensing*, 5–9 November 2001, Singapore, pp. 526–531.

MUSTARA, H., 2006, Pemodelan spasial kerawanan kebakaran di lahan gambut: studi kasus Kabupaten Bengkalis, Provinsi Riau. Master thesis, Bogor Agricultural University, Indonesia (in Bahasa).

PANG-NING, T., MICHAEL, S. and VIPIN, K., 2006, *Introduction to Data Mining* (Saddle River, NJ: Pearson Education, Inc.).

SENG, C.T., WYNNE, H., KIM, H.L. and LEE, C.Y., 2003, Spatial data mining: clustering of hot spots and pattern recognition. In Geoscience and Remote Sensing Symposium, 21–25 July 2003, (New York: IEEE), pp. 3685–3687.

STEPHEN, M., 2009, *Machine Learning: An algorithmic perspective* (Boca Raton, FL: CRC Press).

TJEN-SIEN, L., WEI-YIN, L. and YU-SHAN, S., 2000, A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning Journal*, 40, pp. 203–228.