Classification Model for Forest Fire Hotspot Occurrences Prediction Using ANFIS Algorithm

A K Wijayanto[1], O Sani[2], N D Kartika[2,3], Y Herdiyeni[4]

[1]Center for Environmental Research, Bogor Agricultural University, Indonesia, Gedung PPLH lt 3, Jl Lingkar Akademik, Kampus IPB Darmaga, Bogor, Indonesia, 16680
[2]Master of Science in Information Technology for Natural Resources Management, Bogor Agricultural University SEAMEO BIOTROP Campus, km 6 Tajur, Bogor, West Java, Indonesia, 16721
[3]The Agency of The Assessment and Application of Technology (BPPT), Jakarta, Indonesia
[4]Computational Intelligence Lab, Department of Computer Science, Bogor Agricultural University, Darmaga Campus, Wing 20 Level V-VI, Bogor, West Java, Indonesia, 16680

E-mail: akwijayanto@apps.ipb.ac.id

Abstract. This study proposed the application of data mining technique namely Adaptive Neuro-Fuzzy inference system (ANFIS) on forest fires hotspot data to develop classification models for hotspots occurrence in Central Kalimantan. Hotspot is a point that is indicated as the location of fires. In this study, hotspot distribution is categorized as true alarm and false alarm. ANFIS is a soft computing method in which a given inputoutput data set is expressed in a fuzzy inference system (FIS). The FIS implements a nonlinear mapping from its input space to the output space. The method of this study classified hotspots as target objects by correlating spatial attributes data using three folds in ANFIS algorithm to obtain the best model. The best result obtained from the 3rd fold provided low error for training (error = 0.0093676) and also low error testing result (error = 0.0093676). Attribute of distance to road is the most determining factor that influences the probability of true and false alarm where the level of human activities in this attribute is higher. This classification model can be used to develop early warning system of forest fire.

1. Introduction

The factors that influence fire occurrence in the boreal forest include the properties of the forest vegetation, weather, and ignition agents [1]. Forest fire mostly occurs in Central Kalimantan Province that has large area of peatland and this type of land makes is more dangerous than other forest fire types because fire in peat penetrates the bottom layer of the soil to form a funnel hole, and then the fire spreads horizontally beneath the surface, with addition that climate variations play an important role in influencing peatland fire vulnerability [2].

In Kalimantan, the number of hotspot during the year of 2002-2006 was in Central Kalimantan Province [3]. The impact of forest fire is not only experienced by Indonesian people but also the people who live in neighbour countries near Indonesia. It can affect health, social, politic and economic sector.

The government-led fire management practice in Central Kalimantan has been focused on short-term fire suppression during the May to September dry season. It depends upon a fire monitoring and warning system, using weather and environmental data from agencies such as the Indonesian National Institute of Space and Aeronautics (LAPAN), Ministry of Environment, ASEAN Specialized Meteorological Centre (ASMC), and the Indonesian Bureau of Meteorology, Climatology and Geophysics (referred to as the Indonesian meteorological service or BMKG). At the provincial level, the Natural Resource Conservation Agency (Badan Konservasi Sumberdaya Alam or BKSDA) uses hot spot data and the Fire Spread Risk Index Map (FSRIM), combined with on-the-ground observations, to assess current fire risk levels across the province. [4]

Previous researches developing forest fire risk model have been conducted by using data mining technique as C4.5 algorithm on Weka toolkit: J48, SimpleCart and NaveBayes. Classical data mining methods can extract patterns easily from spatial datasets which is more difficult by other methods, numeric and categorical spatial data types are complex but classical data mining methods do not support locations of objects or relationships between objects that implicitly exist in a spatial dataset [5]. Therefore the methods cannot be utilized to discover knowledge from spatial datasets. Locations of objects determine relations of the objects to its neighbours. So by generating data, classification can be done by data mining algorithm.

Vega-Garcia et al [6] adopted Neural Networks (NN) to predict human-caused wildre occurrence. The data were analyzed using logistic regression analysis (binary logit model), which served as the domain expert to identify the important input variables. The resultant model had four input nodes and two output nodes, and correctly predicted 85 percents of the no-fire observations and 78 percents of the fire observations. Cortez [7] also done a similar research by using data mining approach to predict forest fires using meteorological data. The Data Mining (DM) approach were explored to predict the burned area of forest res. Five different DM techniques, e.g. Support Vector Machines (SVM) and Random Forests, and four distinct feature selection setups (using spatial, temporal, FWI components and weather attributes), were tested on recent real-world data collected from the northeast region of Portugal.

In this study, Adaptive Neuro-Fuzzy Inference System (ANFIS) which is classification algorithms of data mining in Matlab software is used to determine classifier rule. It will show that hotspots occurrences can be predicted by using nearest distance relation between target (hotspot data) and factor (spatial data).

The objective of this project is to develop a decision support system using Adaptive Neuro-Fuzzy Inference System (ANFIS) algorithm for predicting hotspot occurrence in Central Kalimantan Province, Indonesia. The result will be based on two classes: true alarm and false alarm.

2. Methodology

2.1. Materials and Tools

The dataset for this project are physical and hotspot data taken from project research of Spatial Model of Land and Forest Fire Risk Index [3]. Explanatory objects are physical data including land cover, roads, rivers, city centre, village and settlement that will classify the target objects into false or true alarm.

The tools used in this study were ArcGIS and Matlab. ArcGIS used to calculate distance of attributes to hotspot data as target object. Then, ANFIS tool in Matlab used to generate rules in model building.

2.2. Study Area

The study area is in Central Kalimantan Province. The chosen spatial data of the boundary area was based on hotspot occurrence in Central Kalimantan province with coordinate on longitude 113^{o}30'E latitude 02^{o}00'S.

2.3. Overall Step Flow

This study conducted through several steps as shown by Figure 1. Data collected in data collecting step is spatial and hotspot data from a research conducted by Samsuri in 2008 [**?**]. The detail of each next step explained in the next sub-chapter.

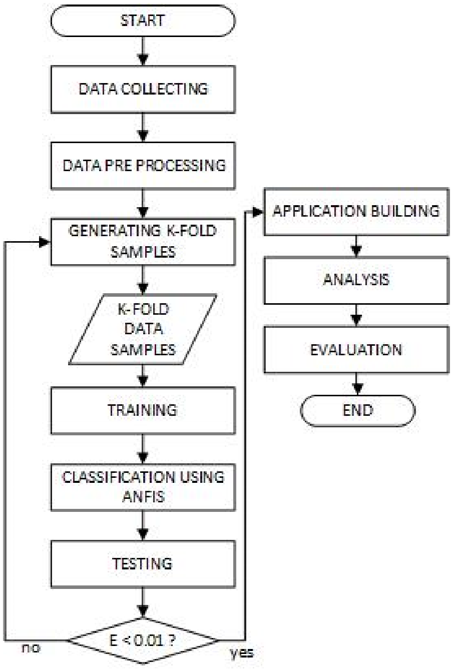

Figure 1. Overall stepflow of research.

2.4. Data Pre-processing

Two main tasks conducted in constructing a forest fires dataset are: 1) creating the target attribute and populating its value from the target objects, and 2) creating explanatory attributes. Target objects are true and false alarm data. True alarm data (positive examples) are hotspots that spread in Central Kalimantan. False alarm data (negative examples) were randomly generated and they are located within the areas at least 1 km away from any true alarm data.

Preprocessing steps were performed to relate the explanatory attributes to target objects by applying topological and metric operations. The relations between target objects and spatial objects which are settlement, roads, rivers, village and city centroids defined by calculating distance from target objects to nearest spatial data by using ArcGIS.

2.5. Generating Rule

Method used to generate rules is hybridization of ANN (Artificial Neural Networks) and FIS (Fuzzy Inference System) which called ANFIS model. The Adaptive Neuro Fuzzy Inference System (ANFIS), was first introduced by Jang, in 1993. The ANFIS is a neural network that is functionally the same as a TakagiSugeno type inference model [8]. The architecture of an ANFIS has introduced the first order Sugeno-style FIS at first. A first-order Sugeno-style FIS model is a system that manages the process of mapping from a given crisp input to a crisp output, using fuzzy set theory [9]. Considering a Sugeno type of fuzzy system having the rule base:

(i) If x is A1 and y is B1, then f1 = c11x+c12y+c10

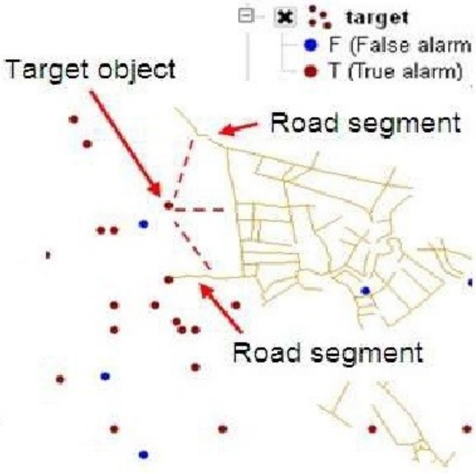

Figure 2. Distance near neighbour between Target object and Road Segment

(ii) If x is A2 and y is B2, then f2 = c21x+c22y+c20

All computations can be presented in a diagram form. ANFIS normally has 5 layers of neurons of which neurons in the same layer are of the same function family (Figure 3)

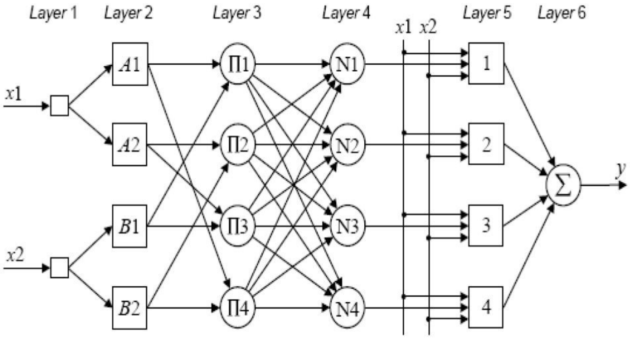

Figure 3. Structure of the ANFIS network

From Figure 3, attribute data (x1) in Layer 1 is divided into three membership functions in Layer 2: near (A1), medium (A2), far (A3). While the next layers become the hidden layers which are structured by ANN algorithm. Layer 3 nodes calculate the ratios of the rules firing strength to the sum of all the rules firing strength. While layer 4 nodes compute a parameter function on the layer 3 output. Parameters in this layer are called consequent parameters. The result is a normalised firing strength. Layer 5 normally a single node that aggregates the overall output as the summation of all incoming signals. The output layer (Layer 6) is the summation layer.

To choose the best combination of training and testing data, cross-validation is done. The basic idea of cross-validation is to split the training set into two disjoint sets, one which is actually used for training, and the other, the validation set, which is used to monitor performance. An extreme case of k-fold cross-validation is obtained for k = n, the number of training cases, also known as leave-one-out cross-validation (LOO-CV). Of- leave-one-out ten the computational cost of LOO-CV (training n models) is prohibitive, but in certain cases, such as Gaussian process regression, there are computational shortcuts [10]

In this study, the data is divided into three k-fold data samples. From the 100 samples data, three samples obtained by combining training and testing data, as shown by Table 1.

Table 1. K-Fold data samples

Fold (disjoint set)	Training-Testing combination
Fold 1	60 tr 40 ts
Fold 2	40 ts 60 tr
Fold 3	20 ts 60 tr - 20 ts

To do the training process, hybrid method is used instead of the backpropagation method by using 100 epochs. In this study, different number (200, 500, and 1000) of epoch are already tried. However, since the error value is already stable on 3rd epoch, so it is not necessary to do too much epoch.

3. Result and Discussion

3.1. Preprocessing Result

The rules were developed from a dataset contained of physical data and target data. The number of training data was 100 and being set randomly. The training data were consisted of 50 true hotspot alarms and 50 false hotspot alarms. The dataset had 6 attributes and each attribute represented:

(a) Distance from target data to nearest city in meter (NEAR_CITY).
(b) Distance from target data to nearest village meter (NEAR_VILLAGE).
(c) Distance from target data to nearest road in meter (NEAR_ROAD).
(d) Distance from target data to nearest river in meter (NEAR_RIVER).
(e) Distance from target data to nearest settlement in meter (NEAR_SETTLEMENT).
(f) Target attribute containing true alarm and false alarm (TARGET)

The distance is calculated by using ArcGIS and stored as a .dat file. This file then being divided into three k-fold samples and used in the training stage. The training stage result is explained in the next sub-chapter.

3.2. Rules Generation

The dataset was split into two disjoint sets consisted of 60 percent of training data and 40 percent of testing data. In this project, the dataset was divided into three samples with different k-fold. All the samples were trained using Matlab anfisedit as a preloaded feature to generate rules for determining true alarm and false alarm and also to get error for each training data.

From Table 2 and Figure 4, the highest error is obtained from fold 1 by combination of first 60 data for training and other 40 data for testing. The lowest score obtained from fold 2 by combination of first 40 data for testing and 60 other data for training. The best fold is not the highest nor the lowest error, but the average. The 3rd fold gives error which is closest to the average. Thus the 3rd fold is chosen to generate rules to determine true and false alarm.

Table 2. Error comparison of 3 folds

Fold (disjoint set)	Training error	Accuracy (a)	a-x̄
Fold 1 (60 tr 40 ts)	0.014777	99.9852	0.0057
Fold 2 (40 ts 60 tr)	0.0030858	99.9969	0.0060
Fold 3 (20 ts 60 tr 20 ts)	0.0093676	99.9906	0.0003
	Average (x̄)	99.9909	

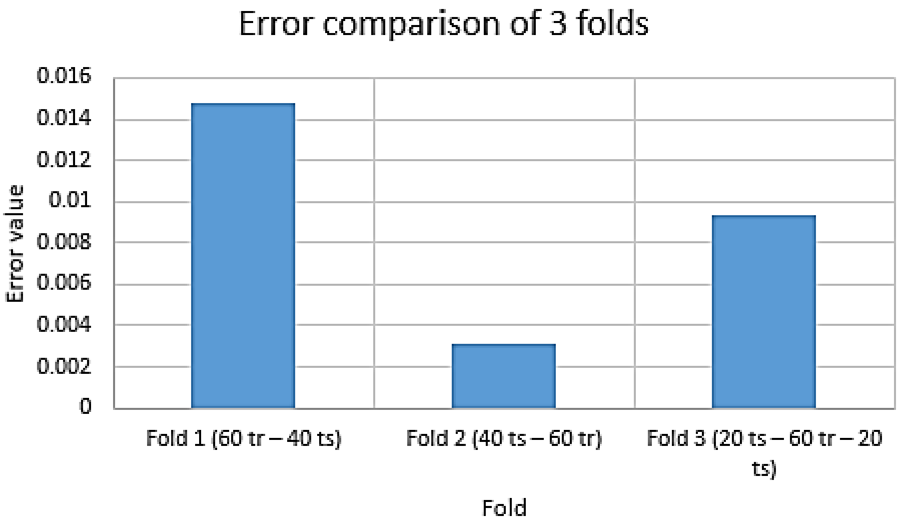

Figure 4. Chart of error comparison of 3 folds

Each input data for ANFIS were divided into three membership functions (near, middle, far) that represented the distance from hotspot to any place where humans did their activities (cities, villages, roads, rivers, settlements). Total node of membership function is fifteen and each node of membership function was generated to be the rule base that represented Sugeno type of fuzzy system function where:

If x is A and B and C and D and E, then Output Parameter

(A,B,C,D, and E is Attribute Input)

The output parameter that is influenced by attribute NEAR_ROAD is the most determined class of true alarm or false alarm. The attribute NEAR_ROAD is the determining factor because the level of human activities that influence forest fire occurrences was high and it means that the probability of hotspot occurrences is also high. This result can be said as match to the result of research conducted by Sitanggang [11] who did a research about hotspot occurrences prediction in Riau Province Indonesia, by using C4.5 algorithm. The research also used attribute of distance to determine the occurrences of forest fire. From the result, it is proven that attribute of distance can be used as determining attribute to forest fire occurrences.

The rule of training result then being tested by using testing data. Testing process provided low error by 0.0093676. Because of the low error, the rule can be used for further utilization.

4. Conclusion and Future Work

ANFIS algorithm can be used to generate expert system for predicting hotspot occurrences. The proposed method provided low error for training result (error = 0.0093676) and also low error testing result (error = 0.0093676). Attribute NEAR_ROAD is the most determine factor that influences the probability of true and false alarm where the level of human activities in this attribute is high. The higher level of human activities, the higher probability of hotspot to be forest fire. This classification model can be used to build early warning system for forest fire.

Another attribute that might determine probability of forest fire occurrences as like land use type, land cover type, and soil type, might be considered in the future work. The more attribute expected to make higher accuracy of prediction.

References

[1] Anderson K, Martell D, Flannigan M and Wang D 2000 *Climate Change, and Carbon Cycling in the Boreal Forest* (New York: Springer)

[2] Syaufina L 2002 *The Effect of Climate Variation on Peat Swamp Forest Condition and Peat Combustibility* Ph.D. thesis University Putra Malaysia

[3] Samsuri 2008 *Spatial Model of Land and Forest Fire Risk Index, Case Study in Central Kalimantan Indonesia* Ph.D. thesis Bogor Agricultural University

[4] Ceccato P, Jaya I, Qian J, Tippet K, Robertson W and Someshwar S 2009 Early warning and response to fires in kalimantan, indonesia *Advances in Operational Weather Systems for Fire Danger Rating* ed Sivakumar M (New York: Springer)

[5] Zeitouni K 2000 *Proc. Information Resources Management Association International Conference on Challenge of Information Technology Management in the 21st century* (Palo Alto)

[6] Vega-Garcia C, Lee B, Woodard P and Titus S 1996 *AI Application* **10** 9

[7] Cortez P and Morasis A 2007 *Proc. the 13th Portuguese Conference on Artificial Intelligence (EPIA 2007), New Trends in Artificial Intelligence* (Guimares)

[8] Jang J 1993 *IEEE Trans On Systems Man and Cybernetics* **19** 665

[9] Loganathan C and Girija K 2013 *International Journal of Engineering and Science* **2** 06

[10] Rasmussen C and Williams C 2006 *Gaussian Processes for Machine Learning* (Massachusetts: The MIT Press)

[11] Sitanggang I, Yakoob R, Mustapha M and Ainuddin A 2012 *Journal of Theoretical and Applied Information Technology* **43** 214