Prediction of Forest Fire using Hybrid SOM-AdaBoost Method

D Rosadi\textsuperscript{1}, W Andriyani\textsuperscript{2}, D Arisanty\textsuperscript{3}, and D Agustina\textsuperscript{4}

\textsuperscript{1}Department of Mathematics, Gadjah Mada University, Indonesia  
\textsuperscript{2}Master of Information Technology, STMIK Akakom, Indonesia  
\textsuperscript{3}Department of Geography Education, Lambung Mangkurat University, Indonesia  
\textsuperscript{4}Department of Statistics, State University of Padang, Indonesia

\textsuperscript{*}E-mail: dedirosadi@gadjahmada.edu

Abstract. Prediction of the occurrences of forest fire has become interest of various research studies for instances, it is found that the hybrid method based on clustering using fuzzy c-means before doing the classification approach will improve the performance of prediction than directly apply the classification approach. In this study, we will consider the new hybrid approach between clustering based on Self Organizing Map (SOM) approach and classification using Boosting (AdaBoost) approach. Our empirical analysis shows using the same public data set, which has been used in several previous studies, the performance of hybrid SOM-AdaBoost will outperforms other methods in literatures.

1. Introduction

It has been discussed in various studies that forest fire prediction is an important step for early warning system in forest fire fighting. The accuracy of prediction of the forest fire events relies heavily on the methods of prediction used in the study. In some recent literature, it is known that various methods can be used to obtain the prediction, namely physics-based model, statistical-mathematics models, and machine learning approach, see e.g. [1].

Recent literature reviews show machine learning approach have become so popular in the study. Using the meteorological and forest weather index (FWI) variables. [2] study the prediction approach based on the Multiple Regression (MR), Decision trees (DT), Random Forests (RF), Neural Networks (NN) and Support Vector Machines (SVM) method. In [3], it was extended the studies in [4-6]) by using proposed hybrid prediction approach based on Fuzzy C-Means clustering and Back-Propagation Neural Networks (BPNN - with one hidden layer) classification. Empirical study in [3] show that this hybrid approach outperforms the classical classification methods for prediction the forest fires. In [7], we extend and show how to improve the performance of approach used in [3] by using the ensemble Adaptive Boosting (AdaBoost) classification approach ([8] and [9]) to replace the BPNN method. In this paper, we consider a hybrid approach based on Kohonen Self-organizing Maps (SOM) clustering which has not been considered in any of the previous studies. For classification, we consider several classical classification approaches, namely SVM, multilogit and DT approach, and ensemble AdaBoost. Here we provide empirical study using the same data as in [2], [3] and [7].

The rest of this paper is presented as follows. In this section we provide the background of the problem. In Section 2, we summarize the necessary theory related to our proposed approach and
provide the proposed algorithms. In Section 3, we discuss the empirical implementation of the algorithms. Section 4 provides the summary of the results.

2. Methods

2.1. Self-organizing maps

Kohonen Self-organizing Maps (SOM) is a special type of neural network model that is configured for data classification. SOM was first introduced by Tuevo Kohonen in 1982 (see e.g., [10]). The Kohonen SOM structure contains only two layers, namely the input and the output layers (see Figure 1).

![Kohonen self-organizing map](image)

Figure 1. Kohonen self-organizing maps (SOM)

SOM does not require a hidden layer because in the process, SOM only maps each neuron in the input layer directly to each neuron in the output layer so that each neuron in the output layer represents a cluster of the given input. It is also known that each output neuron receives input through a weight that is directly connected to the input so that the weight vector has the same dimensions as the input vector. SOM uses unsupervised learning techniques and is an efficient technique for clustering data that has very large dimensions into low-dimensional data.

There are 4 steps to classify data using SOM (Self Organizing Map):

1. Competition: for each of the output node \( j \), determine the distance between the vector of input node \( x \) and weight vector of the \( m \)-th neuron node \( w_i \). It can be expressed as \( D(x, w_i) \), is calculated using Euclidean distance function, which is defined as

\[
d(x, w_m) = \sqrt{\sum_{i=1}^{n} (x_i - w_{mi})^2}
\]

2. Updated weight: winning neuron select from the minimum distance from each weight vector and input vector. Adaption processes is applied for each winning neuron and their neighbourhood by updating the weight value where \( f(n) \) denotes the node function and \( n \) is the iterations number.

The Gauss function can be used as the node neighbourhood function as follow:

\[
n(t) = a(n) \times e^{-\frac{||r_x-r_y||^2}{2\sigma^2(n)}}
\]

Where \( a(n) \) is alpha value (learning rate). Learning rate is a function of decreasing learning rate over time. Here \( ||r_x-r_y||^2 \) defines the squared distance between the \( x \)-neuron and the winning neuron in the grid and \( \sigma(n) \) is the width of the neighbourhood used. Learning rate value is obtained from:

\[
a(n) = a_i \left( 1 - \frac{n}{n_{max}} \right)
\]
where $\alpha_i$ is the initial learning rate and $n_{\text{max}}$ is the maximum number of iteration. The formula of width of the neighbourhood i.e:

$$\sigma(n) = \alpha_i \left( \frac{\sigma_m}{\sigma_i} \right)^{\frac{n}{n_{\text{max}}}}$$

$\sigma(n)$ is the width of the neighbourhood. It will decrease with increase of $n$ learning steps. Here $\sigma_i$ denotes the initial of neighbourhood width and $\sigma_m$ defines the final value of neighborhood width.

3. Updates $\alpha$ and $\sigma$ with the formula in step 2.

4. Stopping criteria. Stop training if the stopping criteria is fulfilled, i.e., when the number of iterations or the minimum error value, the alpha value, and the width of the neighbours reached.

2.2. Classical classification methods

To compare the performance of the method, we apply several well-known classification approaches to the data such as radial-support vector machine (SVM), multinomial Logistic Regression (multilog) and Decision Tree (DT) method. A short summary of the methods are available in, e.g., [11]. See e.g. [12], [13], [14], [15] and [16] for further detail.

2.3. Ensemble classification: adaboost method

Adaptive Boosting (AdaBoost), is an ensemble classification approach. Using AdaBoost, the ‘weak learners’ is combined with the other boost classifier, which has the form

$$F_T(x) = \sum_{t=1}^{T} f_t(x)$$

Here each $f$ denotes a ‘weak learner’ that uses an object $x$ as input and provides the output as the class of the input $x$. There are various algorithms for implementing AdaBoost method, here we apply what so called Breiman-AdaBoost.M1 algorithm. See [9] for the description detail.

3. Results and discussion

3.1. Data description

Here we use the same data as the study in [2],[3] and [7] and it may be obtained from UCI machine learning websites. The original dataset has 12 variables, which contain the meteorological and forest weather index (FWI) variables. The original dataset contains 517 cases, from year 2000 until 2007. For the study here, however, we only use FFMC, DMC, DC, ISI, Temperature, RH, Wind, Rain and area variables.

3.2. Algoritms

Because we extend the study in [7], we have to follow closely the algorithm uses in [7] with some necessary updated, as follows

Preprocessing steps

1. First, we do preprocess by splitting the data into two class labelled as “No Burned Area” (whenever the variable area is 0) and “Burned Area” (whenever the variable area is greater than 0)

2. We apply normalization to the variables for “No Burned Area” data only. Here for instance, as in our previous study, we apply min-max normalization, which is defined as

$$v_i = \frac{v_i - \min A}{\max A}$$

Here $\min A$ and $\max A$ denote the minimum and the maximum values of a variable $A$, $v_i$ denotes the data value in attribute $A$ that will be mapped into the new range.
Clustering step
3. As the new clustering approach, in this study, we apply SOM clustering into “Burned Area” data to obtain “Light Burned Area” and the “Heavy Burned Area” cluster. Here we apply command `som` in package `kohonen` [17] in R [18].
4. For “Burned Area” data only, apply min-max normalization to the variables, and combine it with the cluster information obtained in step 3.
5. Combine data of “No Burned Area” and the data obtained from step 4 to obtain three classes of the data.

Classification step
6. For obtaining the testing and training data, we randomly split the data into training data (70 percent, 80 percent, and 90 percent) and testing data (30 percent, 20 percent, and 10 percent).
7. Implement the classification algorithm on the training data using the considered approaches.
8. As the numerical performance evaluation, we consider the accuracy measure, defined as

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

where \(TP\), \(TN\), \(FP\), and \(FN\) defines the true positive, the true negative, the false positive and the false negative cases, respectively, in the categorical classification data. Other numerical performance measures are also available, however, to save space, we will not report these results in the empirical study.

For comparison purpose, we also apply the fuzzy c-mean clustering in step 3 above, using the command `fcm` in the package `ppclust` [19]. For classification step, we apply the command `multinom` in the package `nnet` [20] for implementation of multilogit algorithms, Decision Tree approach using function `rpart` in package `rpart` [21], SVM classification approach using `svm` command in package `e1071` [22], and function `boosting` in the package `adabag` [23] in R for implementation of Breiman-AdaBoost [9] algorithms, respectively.

**Figure 2.** SOM cluster map
Table 1. Summary of the performance of hybrid methods

| Data Training: | Fuzzy C-Means clustering using cosine distance similarity | SOM-kmeans clustering |
|---------------|--------------------------------------------------------|-----------------------|
|               | Decision Tree | Breiman-AdaBoost | Multi-logit | SVM | Decision Tree | Breiman-AdaBoost | Multi-logit | SVM |
| Data Testing  | Accuracy     | Accuracy         | Accuracy     | Accuracy | Accuracy     | Accuracy         | Accuracy     | Accuracy |
| 70%           | 0.9503       | 1                | 0.9475       | 0.9972  | 0.9779       | 1                | 0.9669       | 0.9972  |
| 30%           | 0.9548       | 0.9871           | 0.9613       | 0.9419  | 0.9871       | 1                | 0.9677       | 0.9548  |
| 80%           | 0.9396       | 1                | 0.9517       | 1       | 0.9710       | 1                | 0.9686       | 0.9976  |
| 20%           | 0.9515       | 0.9806           | 0.9806       | 0.9612  | 0.9903       | 1                | 0.9903       | 0.9709  |
| 90%           | 0.9462       | 1                | 0.9505       | 1       | 0.9742       | 1                | 0.9806       | 0.9978  |
| 10%           | 0.9231       | 0.9808           | 0.9615       | 0.9231  | 0.9808       | 1                | 0.9808       | 0.9615  |

3.3. Discussion
Noted that for clustering the SOM data space, we apply hierarchical clustering (hclust) of the distance between SOM grid and compare it with the performance of k-mean clustering over the SOM grid space. The clusters for the grid space are given in Figure 2. For checking the performance of considered hybrid clustering and classification methods, we consider several trainings and testing sample sizes. The numerical summary of performance of each method in data training and data testing are given in Table 1. The performance of SOM-kmeans clustering is better than fuzzy-c-mean approach, where the Breiman-AdaBoost approach will be the best either in sample or out sample data. Therefore, in general, this empirical study shows that the hybrid algorithms between SOM clustering approach with AdaBoost Classification method will be a promising approach to predict the size of the forest fire.

4. Conclusion
It has been discussed in various studies that forest fire prediction is an important step for early warning system in forest fire fighting. In this paper, we study the prediction of the forest fire size occurrences using meteorological and FWI variables. This study shows that the hybrid algorithms between the fuzzy c-means clustering or SOM classification approach with AdaBoost Classification method will be promising approaches to predict the size of the forest fire. It is also outperform of the other considered method in the literature. We suggest the interested reader to further study other hybrid approaches as the alternatives of the method considered in this paper.

Acknowledgement
The authors acknowledge the research funding PD 2021 from Deputy of Research and Development, National Agency for Research and Innovation, Ministry of Education and Culture, Republic of Indonesia.

References
[1] Bui D T, Bui Q T, Nguyen Q P, Pradhan B, Nampak H and Trinh P T 2017 Agric. For. Meteorol. 233 32
[2] Cortez P and Morais A 2007 Proc 13th EPIA 1 512
[3] Shidik G F and Mustofa K 2014 Inf. Commun. Technol.-EURAsia Conf. (Heidelberg: Springer) p 316
[4] Ganesan P, Sathish B S and Sajiv G 2016 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave) p 1
[5] Wang Z, Liu P, and Cui T 2017 *Int. Arch. Photogramm. Rem. Sens. Spat. Inf. Sci.* 42 925
[6] Hidayati I C, Nalaratih N, Shabrina A, Wahyun I N and Latifah A L 2020 *For. Soc.* 4 280
[7] Rosadi D and Andriyani W 2021 *Journal of Physics: Conf. Series ICMSE 2020* 1918
[8] Zhu J, Zou H, Rosset S and Hastie T 2009 *Stat. Interface* 2(3) 349
[9] Freund Y and Schapire R E 1996 *Proc. 13th Int. Conf. on ML* (Massachusetts :Morgan Kaufmann) p 148
[10] Kohonen T 1989 *Self-organizing feature maps. Self-organization and associative memory.* (Heidelberg: Springer)
[11] Rosadi D, Andriyani W, Arisanty D and Agustina D 2021 *IEEE Proc. 2020 3rd ISRITI*
[12] Hastie T, Tibshirani R and Friedman J 2009 *The Elements of Statistical Learning* (New York: Springer)
[13] Christmann A and Steinwart I 2008 *Support Vector Machines* (New York: Springer)
[14] Cichosz P 2015 *Data Mining Algorithms: Explained Using R* (NY: John Wiley & Sons)
[15] James G, Witten D, Hastie T and Tibshirani R 2013 *An Introduction to Statistical Learning: with Applications in R* (NY: Springer)
[16] Rokach L and Maimoon O 2008 *Data Mining with Decision Trees: Theory and Applications*, (Singapore: World Scientific)
[17] Wehrens R, Kruisselbrink J 2018 *J of Stat. Soft.* 87 1
[18] R CT 2021 *R Foundation for Statistical Computing* (Vienna: RCTeam)
[19] Cebeci Z 2019 *Journal of Agricultural Inform* 10 1
[20] Venables W N, Ripley B D 2002 *Modern Applied Statistics with S* (New York: Springer)
[21] Therneau T, Atkinson B and Ripley B 2019 *User Guide of Package `rpart`* (Vienna: CRAN)
[22] Meyer D 2021 *User Guide of Package `e1071`* (Vienna: CRAN)
[23] Alfaro E, Gamez M.and Garcia N. 2013 *Journal of Stat. Soft.*, 54 1