Prediction of forest fire using ensemble method

D Rosadi, W. Andriyani

1Department of Mathematics, Gadjah Mada University, Indonesia
2STMIK Akakom, Indonesia

*Corresponding author: dedirosadi@gadjahmada.edu

Abstract. In this paper we consider the application of ensemble classification method, which is called as the Adaptive Boosting (AdaBoost) method, to predict the occurrences of forest fire. To illustrate the method, we consider the application of the method using the same public data set, which has been used in the previous studies, but the ensemble approach is not considered in these studies yet. We also compare the performance of the ensemble method with several other classical classification methods, such as the Decision tree and SVM method. All computation are done using open source software R. We find that in the empirical study, the hybrid algorithms between the fuzzy c-means clustering and the ensemble approach will outperform the other classification methods considered in the study.

1. Introduction

Forest fire is an important environmental world phenomenon and has many impacts to human life, infrastructures, and the environment. One of the key successes that help the forest fire firefighting is the early warning systems of fire detection. This system is related to accurate prediction of fire condition which can be caused by several variables. For forest fire prediction, there are various techniques in literature has been proposed which can be classified into three main models, namely physics-based model, statistical models, and machine learning models, see e.g. [4]. It is also known that the data and variables that often used in the literature are coming from various sources, such as satellite data (infra-red or hotspots sensors) and ground sensors (e.g., the weather and or other meteorological data), see e.g. [1].

Among many approaches, machine learning and data mining approach has receive many attention by many researchers. Based on meteorological and forest weather index (FWI) variables. [1] studied classification algorithm, namely the Multiple Regression (MR), Decision trees (DT), Random Forests (RF), Neural Networks (NN) and Support Vector Machines (SVM) to model the forest fire prediction. More related recent studies are available. Wijayanto [5] applied the classification algorithm for hotspot occurrence using Adaptive Neuro-Fuzzy inference system (ANFIS) system. Vega-Garcia et al [6] considered Neural Networks (NN) and logistic regression analysis (binary logit model) for forest prediction. In particular, [2] considers hybrid approach between clustering technique, normalization to preprocess the data and apply classification approach to the normalized data. They show that the combination of the Fuzzy C-Means clustering with Cosine distance, Min-Max normalization and Back-Propagation Neural Networks (with one hidden layer) classification method can give a relatively accurate prediction compare to other classification approach (SVM, K-Nearest Neighborhood, DT, and Naive Bayes) and the case without the clustering the data. It is known in literature that k-means and/or Fuzzy C-means (FCM) clustering methods has been used to cluster the fires phenomenon (see e.g. [7-9]). In this paper, we improve the performance of method considered in [2] by considering the
ensemble method, in particular applying what so called the Adaptive Boosting (AdaBoost) approach [10]. We also check the performance of method by randomly splitting the data into data training and data testing, to obtain more objective measure of performance of the considered method. In the empirical study, we compare the performance of AdaBoost with the performance of DT and SVM method.

The rest of this paper is organized as follows. In this section we already provide a quick introduction to the problem that we considered in this paper. In Section 2, we outline short description of necessary theory related to our considered approach and provide the algorithms. In Section 3, we provide empirical results. Section 4 concludes the results.

2. Methods

2.1. Clustering Method: Fuzzy C- Means Clustering
Fuzzy C-Means (FCM) clustering algorithm was proposed by [11,12] and can be considered as the generalization of K-means clustering approach. Given a finite set of data \( X = \{x_1, \cdots, x_n\} \), the FCM algorithm will classify the data into \( m \) fuzzy clusters. Here the algorithms will give the \( m \) cluster centers \( C = \{c_1, \cdots, c_m\} \) and a partition matrix \( W = w_{ij} \in [0,1], i = 1, \cdots, n, j = 1, \cdots, m \) where each element of \( W \), i.e. \( w_{ij} \) gives the membership degree of each element \( x_i \) belongs to cluster \( c_j \). See e.g. [13] for further detail.

2.2. Ensemble Classification: Adaboost Method
Adaptive Boosting or AdaBoost, is a machine learning ensemble approach. Using AdaBoost, the performance of the 'weak learners' can be improved by combining them into a weighted sum that is called as a boost classifier, which has the form
\[
F_t(x) = \sum_{i=1}^{T} f_i(x)
\]
where each \( f \) is a weak learner that takes an object \( x \) as input and returns the class of the input \( x \). There are various variations of AdaBoost algorithms, the algorithm used in this study is called Stagewise Additive Modeling using a Multi-class Exponential loss function (SAMME) AdaBoost. See [10] for the detail.

3. Results and Discussion

3.1. Data Description
For the empirical study, we use the same data as the study in [1] and [2], which is available in the UCI machine learning repository. The dataset contains 12 variables, which is describing the meteorological and forest weather index (FWI) variables. The dataset has a total of 517 samples, from year 2000 until 2007. Following [2], here we only use 8 variables, namely FFMC, DMC, DC, ISI, Temperature, RH, Wind and Rain.

3.2. Algorithms
The algorithms we use following closely with [2], with some improvements, as follows

**Preprocessing steps**
1. We use all of the data and split the data into two categories, which is the case of data with the variable area has the value 0 and labeled as “No Burned Area” and the case of data area has the value larger than 0 and labeled as “Burned Area”
2. Normalize all of the seven variables in both of the categories using min-max normalization, which is defined as
\[ v'_i = \frac{v_i - \min A}{\max A} \]

where \( \min A \) and \( \max A \) is the minimum and the maximum values of an variable \( A \), \( v_i \) denotes the data value in attribute \( A \) that will be mapped into the new range.

**Clustering step**

3. We cluster the “Burned Area” data using fuzzy c-means approach into two classes, namely the “Light Burned Area” and the “Heavy Burned Area”. Here we apply the command `fcm` in package `ppclust` in R. For the distance similarity measurement, various approaches have been implemented in R, we only consider to use two approaches, which is cosine and correlation distance metric.

4. We combine the data of “No Burned Area” and the clustered data

**Classification step**

5. We randomly split the data into training data (70 percent and 80 percent) and testing data (30 percent and 20 percent)

6. We apply the AdaBoost approach to the training data and the best obtain model is tested in to the testing data. Here we use the command `boosting` in the package `adabag` in R for implementation of SAMME [10] algorithms.

7. To check the performance of the classification method, we use the accuracy measure, defined as

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

where \( TP, TN, FP, FN \) denotes the true positive, the true negative, the false positive and the false negative cases, respectively, in the categorical classification data. An additional measure is the Cohen’s Kappa statistic measurement [14].

For comparison purpose, we also implement the DT and SVM classification approach in step 6. It is implemented in R using function `rpart` in package `rpart` [15] and `svm` in package `e1071` [16], respectively.

**3.3. Discussion**

The summary of the empirical results is given in Table 1. Here we consider several training and testing sample sizes for checking the performance of considered clustering and classification methods. The performance of each method in data training and data testing are given in the table. In can be seen that the SVM method will perform the worst either for classify the training or the testing data. In the training data, Adaboost method will always improve the performance of DT, which is the weaker version of the classification method. In the data testing, Adaboost seems to be most of the time will perform the best to classify the data testing. The performance of Adaboost sometimes can reach 100 percent accuracy for our considered study. We also may see that the cosine distance metric will be always better for the data training, but the results mixed for the data testing. Therefore, in general, this empirical study show that the hybrid algorithms between the fuzzy c-means clustering and the Adaboost ensemble will be a promising approach to predict the size of the forest fire.
### Table 1. Summary of the performance of hybrid methods

| Fuzzy C-Means (distance similarity) | Data Training | Data Testing | SAMME-AdaBoost | DT | SVM |
|-------------------------------------|---------------|--------------|----------------|----|-----|
| Cosine                              | 70%           | 30%          | 0.9862         | 0.9758 | 0.9696 | 0.9458 |
| Correlation                         | 70%           | 30%          | 0.9806         | 0.9687 | 0.8581 | 0.7638 |
| Cosine                              | 80%           | 20%          | 0.9879         | 0.9789 | 0.971 | 0.949 |
| Correlation                         | 80%           | 20%          | 0.9709         | 0.9527 | 0.9686 | 0.8738 | 0.7905 |
| Cosine                              | 90%           | 10%          | 0.9849         | 0.9741 | 0.9505 | 0.9136 |
| Correlation                         | 90%           | 10%          | 0.9849         | 0.9741 | 0.9505 | 0.9136 |

### 4. Conclusion

One of the key successes that help the forest fire firefighting is the early warning systems of fire detection. The study of prediction of the forest fire size occurrences using meteorological and FWI variables is one of the important study in early warning fire detection. This study show that the hybrid algorithms between the fuzzy c-means clustering and the Adaboost ensemble will be a promising approach to predict the size of the forest fire and outperform of the other considered method in the literature. We suggest the interested reader to further study other hybrid approaches between the clustering methods and the boosting or ensemble methods as the alternatives of the approach considered in this paper.

### Acknowledgement

The authors acknowledge the research funding PD 2020 from Deputy of Research and Development, National Agency for Research and Innovation, Ministry of Education and Culture, Republic of Indonesia.

### References

[1] Cortez P and Morais A 2007 Proc 13th EPIA 1 512
[2] Shidik G F and Mustofa K 2014 Inf. Commun. Technol.-EurAsia Conf. (Berlin, Heidelberg: Springer) p 316
[3] R CT 2020 R Foundation for Statistical Computing (Vienna: RCTeam)
[4] Bui D T, Bui Q T, Nguyen Q P, Pradhan B, Nampak H and Trinh P T 2017 Agric. For. Meteorol. 233 32
[5] Wijayanto A K, Sani O, Kartika N D and Herdiyeni Y 2017 IOP Conf. Ser.: Earth Environ. Sci. 54(1) 012059
[6] Vega-Garcia C, Lee B, Woodard P and Titus S 1996 AI Application 10 9
[7] 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
[8] Wang Z, Liu P, and Cui T 2017 Int. Arch. Photogramm. Rem. Sens. Spat. Inf. Sci. 42 925
[9] Hidayati I C, Nalaratih N, Shabrina A, Wahyun I N and Latifah A L 2020 For. Soc. 4 280
[10] Zhu J, Zou H, Rosset S and Hastie T 2009 Stat. Interface 2(3) 349
[11] Dunn J C 1973 J. Cybernetics 3 32
[12] Bezdek J C 1981 Pattern recognition with fuzzy objective function algorithms (New York:
Plenum)

[13] Chattopadhyay S, Pratihar D K and Sarkar S C D 2011 *Comput. Inform.* **30** 701
[14] Byrt T, Bishop J and Carlin J B 1993 *J. Clin. Epidemiol.* **46** 423
[15] Therneau T, Atkinson, B and Ripley B 2019 *User Guide of Package ‘rpart’* (Vienna: CRAN)
[16] Meyer D 2020 *User Guide of Package ‘e1071’* (Vienna: CRAN)