



Forest fire forecasting using ensemble learning approaches

Ying Xie¹ · Minggang Peng¹

Received: 27 December 2017 / Accepted: 27 April 2018 / Published online: 11 May 2018
© The Natural Computing Applications Forum 2018

Abstract

Frequent and intense forest fires have posed severe challenges to forest management in many countries worldwide. Since human experts may overlook important signals, the development of reliable prediction models with various types of data generated by automatic tools is crucial for establishing rigorous and effective forest firefighting plans. This study applied recently emerged ensemble learning methods to predict the burned area of forest fires and the occurrence of large-scale forest fires using the forest fire dataset from the University of California, Irvine machine learning repository collected from the northeastern region of Portugal. The results showed that the tuned random forest approach performed better than other regression models did with regard to the prediction accuracy of the burned area. In addition, extreme gradient boosting outperformed other classification models in terms of its predictive accuracy of large-scale fire occurrences. The findings showed that ensemble learning methods not only have great potential for broader application in forest fire automatic precaution and prevention systems but also provide important techniques for forest firefighting decision making in terms of fire resource allocation and strategies, which can ultimately improve the efficiency of forest fire management worldwide.

Keywords Ensemble learning · Random forests · Extreme gradient boosting · Large-scale forest fires

1 Introduction

Forests, which account for more than 31% of the world's land surface [1], contribute to the continuity of ecological balance and play a paramount role in environmental sustainability. As one of the major threats to forest preservation, forest fires have created immeasurable economic and ecological damages and have resulted in enormous human suffering. Each year, millions of hectares of forests are destroyed from various fires, thus consuming a large fraction of firefighting expenses around the world. However, these increasing expenses do not guarantee success in controlling this threat. According to the data collected by World Fire [2], an average of 3.8 million fire incidents occurred per year from 1993 to 2014, and more than 0.9

million human inhabitants died because of wildfires during this period. Forest fires can be caused by a variety of factors, such as lightning, human negligence, rockfall sparks, spontaneous combustion and volcanic eruptions [3–5]. In addition, the causes of forest fires vary throughout the world. The severe impacts of fires on forests have made it imperative for decision makers to identify efficient ways to contain this threat.

Consequently, understanding the factors that influence the occurrence of forest fires is crucial and necessary for the resource allocation of fire prevention, fire suppression and forest management. In addition, the ability to predict fire progression and burned areas is of great importance to mitigate the disastrous consequences of forest fires. A wide range of automatic detection and prediction techniques has been developed to serve this purpose [6]. Given that traditional human surveillance is expensive and may be subject to cognitive limitations, automatic tools, such as satellite-based tools, infrared/smoke scanners and local meteorological sensors, have been developed to monitor and fight forest fires [7, 8]. Over the past decade, classical statistical analysis has given way to several data mining

✉ Minggang Peng
pengminggang@gzhu.edu.cn

Ying Xie
xysoc@gzhu.edu.cn

¹ School of Public Administration, Guangzhou University, Guangzhou 510006, Guangdong, China

techniques in the domain of fire prevention and precaution due to the better accessibility of various types of data, as facilitated by automatic tools [9]. Since human experts are limited in number and may ignore important signals, interest in machine learning has grown since such methods can be used to determine the driving factors and the occurrence probability of forest fires that are triggered by multiple causes; these methods are also expected to improve the accuracy and efficiency of decision making with respect to fire management and preventions [6, 9–15]. Increasingly sophisticated data mining techniques have helped decision makers accommodate large amounts of data in a timelier manner.

Indeed, previous studies have proposed multiple data mining techniques to forecast the spatial distribution of wildfire occurrences or ignitions, including regression trees (RTs) [16], artificial neural networks (ANNs) [7, 17, 18], support vector machines (SVMs) [6, 19] and random forests (RFs) [20]. Driven by advancements in the field of statistics, ensemble learning and deep learning (DL) methods have become major tools in machine learning and have gained tremendous achievements [21]. However, the full potential of ensemble learning methods has yet to be explored, particularly in many fields of decision making, such as natural resource management and wildfire occurrences [22].

Therefore, our objective with this study is to explore and evaluate the potential of ensemble learning methods in greater depth to allow accurate forecasting of the burned area of forest fires and the occurrences of large-scale forest fires by comparing newly developed methods to other stochastic and deterministic data mining techniques and assessing their applications in forest fire forecasting. Specifically, two experiments using ensemble learning methods and other modeling methods were conducted, and their results were compared and calculated to evaluate their performance. The first and second experiments tested the predictive performance in forecasting the burned area and the predictive accuracy in forecasting the occurrence of large-scales forest fires, respectively. Each method was applied to the widely used forest fire dataset (FFDS) from the UCI machine learning repository collected from the northeastern region of Portugal, and the results were compared to assess the prediction accuracy and performance. By evaluating various regression and classification methods, this study determined whether recently emerged ensemble learning methods could provide better forest fire predictions. We expect these findings to be particularly useful in fire management decision making and resource planning.

This paper is organized as follows. Section 2 reviews several machine learning approaches in forest fire management, and Sect. 3 introduces the ensemble learning

methods, including RFs and extreme gradient boosting (EGB). Section 4 describes the study area and study data, and Sect. 5 presents the results of the two experiments. Finally, conclusions are drawn in Sect. 6.

2 Machine learning approaches in forest fire management

Using a series of meteorological indicators, traditional approaches were conducted based on linear or logistic regressions to rate wildfire risks in a relatively short period. However, due to the variances of terrain features, the impacts of various factors on forest fires are not always identical, and the frequencies of wildfire are not significantly related to local temperature, thus resulting in low accuracy when forecasting wildfires. In recent years, forest fire modeling has attracted broad attention, and the several models available now include a series of anthropogenic and meteorological components in their assessments [23, 24]. Machine learning models have demonstrated their accuracy in data mining and other approaches. Thus, a plethora of machine learning algorithms exists to model the spatial distribution of forest fire occurrences or ignitions, including RTs, ANNs, SVMs and RFs.

The RT is an approach used in wildfire risk assessment. In a study of fire-prone areas in Southeast Italy, Amatulli et al. [16] developed the CART analysis to highlight the hierarchical relationships among the predictor variables, in which the improved interpretability of the regression rules represented a tool that was possibly useful for the assessment and representation of fire risks. However, this machine learning approach is not entirely robust because each division involves a set of variables with similar discriminatory power [25]. Therefore, small changes in the data may produce different models. To solve these problems, scholars in the field of data mining have developed ensemble learning methods that generate multiple classifiers and enable the grouping of the results in a final classification that includes boosting and bagging.

The SVM has been the most commonly used method in the detection and prediction of forest fires. For the detection of wildfires, the threshold should be defined, and specific parameters such as temperature and relative humidity should be well predefined. If the threshold value does not correspond with the sensor reading, alarms are triggered. The SVM can be applied at the base station with the polynomial kernel function, and the SVM algorithm can make forest fire predictions even at the risk of generating some mistakes. The SVM attempts to find the optimal hyperplane of separation between classes. The examples located on this hyperplane are called support vectors, which are the most challenging to classify for their lower

separability. In the simplest cases, the optimal hyperplane is defined using a straight line, and the data are linearly separable. In the study of North American forests, an SVM that is fed satellite images can obtain 75% accuracy at finding smoke at the 1.1-km pixel level [19]. Using an SVM and four distinct attribute selection setups, Cortez and Morais confirmed that the SVM was better able to predict small fires, which account for the majority of fire occurrences, than were four other data mining techniques [6]. However, even though the SVM may yield an AUC value higher than that from the logistic regression in the studies, Rodrigues considered this method inadequate for classifying wildfire occurrences since its calibration is extremely time-consuming [25].

Other machine learning methods have also been used to detect wildfires. For instance, Arrue et al. [7] reduced fire false alarms with 90% accuracy in combination with infrared scanners and a neural network (NN). Based on spatial clustering using FASTCiD, Hsu et al. [26] detected forest fire spots in satellite images. Using satellite-based and meteorological data, Stojanova et al. [14] confirmed that a bagging decision tree (DT) demonstrated a high prediction accuracy (with an overall 80% accuracy) of fire occurrences in Slovenian forests.

3 Ensemble learning methods

Ensemble learning is a machine learning paradigm in which multiple learners, such as classifiers or regressions, are strategically assembled to solve a particular statistical problem with the aim of improving classification and regression results [22, 27]. An ensemble includes a number of base learners [28], and the overall generalization ability of an ensemble is typically much stronger than that of individual base learners. Thus, the technique of ensemble learning can increase the accuracy of predictions for the weak learners. RFs and EGB, two of the most popular ensemble learning methods, were adopted in the present study to construct the classification and regression models. Their accuracies were tested against various other linear or DL classification and regression models.

3.1 Random forests

The RF is an ensemble learning method for classification and regression that operates by constructing numerous DTs and producing the best result of classification or prediction (regression) based on the combination of individual trees. Random decision forests can modify the habits of DTs regarding the over-fitting to their training set [29–31]. RFs have a high prediction accuracy and can tolerate outliers and noise [30]. In addition, RFs can select important

variables and identify the relative importance of each independent variable automatically [29, 32]. Due to the strengths of RFs, an increasing number of studies regarding fire occurrences have employed this method to make predictions [33, 34].

The general techniques of bootstrap aggregation or bagging are applied in the training procedure for RFs. The following is the description of RFs that is executed in the Python sklearn classification method.

1. Given a training set $X = \{x_1, \dots, x_n\}$ with their responses $Y = \{y_1, \dots, y_n\}$, random samples are selected (B times) with replacement from the training set and are used to train DTs.

For $b = 1, \dots, B$:

- (a) Sample B training examples from $\{X, Y\}$ with replacement, call these $\{X_b, Y_b\}$, and
- (b) Train an RT f_b on $\{X_b, Y_b\}$.

2. After training, the prediction for the unseen sample x' can be made by averaging the predictions from all the trained individual RTs on x' as

$$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x')$$

or by taking the majority vote in the case of DTs.

The RF is a nonparametric modeling method that can reduce variance and bias. The averaging process over multiple trees can notably reduce instability. Since at least several opportunities exist for a predictor of an individual tree to be the predictor that defines a split, the gains from averaging over a large number of trees (variance reduction) can be significant.

3.2 Extreme gradient boosting

Recently, considerable attention has been paid to the EGB algorithm. Gradient boosting produces a prediction model in the form of an ensemble of weak prediction models, which are typically DTs, for regression and classification problems [35]. By optimizing an arbitrary differentiable loss function on a suitable cost function [31], EGB forms a tree ensemble model constructed by a set of classification and RTs that attempt to define and optimize an objective function. The most important features, such as the gain, cover and frequency, are ordered in the EGB algorithm. The gain provides an indication of how important a feature is in making a purer branch of a DT. The cover measures the relative quantity of observations that are concerned by a feature. The frequency counts the number of times a feature is used in all the generated trees [36]. The present study uses the gradient boosting regression tree (GBRT),

which considers the additive models and is described as follows:

$$F(x) = \sum_{m=1}^M \gamma_m h_m(x)$$

where $h_m(x)$ are the basis functions that are typically called weak learners in the context of boosting. The GBRT uses DTs of fixed size as weak learners. DTs have a number of abilities that make them valuable for boosting—specifically, the ability to accommodate data of mixed types and to model complex functions.

Similar to other boosting algorithms, the GBRT builds the additive model in a forward fashion as follows:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

At each stage, the DT $h_m(x)$ is chosen to minimize the loss function L based on the current model F_{m-1} and its fit $F_{m-1}(x_i)$.

$$F_m(x) = F_{m-1}(x) + \min_h \sum_{i=1}^n L(y_i, F_{m-1}(x_i) - h(x))$$

The initial model F_0 is problem specific. For example, one typically chooses the mean of the target values for the least-squares regression.

Gradient boosting attempts to solve this minimization problem numerically via the steepest descent, of which the steepest direction is the negative gradient of the loss function evaluated at the current model F_{m-1} . This function can be calculated for any differentiable loss function as

$$F_m(x) = F_{m-1}(x) + \gamma_m \sum_{i=1}^n \nabla_F L(y_i, F_{m-1}(x_i))$$

where the step length γ_m is chosen using the line search below.

$$\gamma_m = \arg \min_h \sum_{i=1}^n L\left(y_i, F_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right)$$

4 Location and data

4.1 Study area

This study uses forest fire data from Montesinho Natural Park located in the municipalities of Vinhais and Bragança in the Trás-os-Montes region of Portugal (Fig. 1). The park, which is part of the biosphere reserve of the Iberian plateau with a smooth and rounded landscape embossment, occupies an area of approximately 74,229 hectares on the border with Spain. The altitude ranges between 438 and 1482 m, with valleys that are separated by rivers.

Dominated by a supra-Mediterranean climate, the average annual temperature ranges from 8 to 12 °C. The park contains a high degree of flora and fauna diversity that arises from a great variety of climatic, topographic, environmental and geomorphological conditions and human activities that have shaped those landscapes for millennia. These conditions have facilitated certain species to spread; highlights include heather bushes, rockroses and brooms, natural meadows, chestnut groves, holm oak groves, riverside ecosystems, ultrabasic vegetation and oak woods. Therefore, the complexity of the geographical conditions and the diversity of biophysical conditions play a determining role and are particularly important when modeling the forest fire occurrences since these conditions can strengthen the findings for improved generalization in our study.

4.2 Data

The FFDS provided by the well-known UCI machine learning repository was employed in the experiments. This repository contains 517 wildfires in Montesinho Natural Park in the Trás-os-Montes region of Portugal (Fig. 1) from January 2000 to December 2004. For each wildfire, 12 attributes were registered on a daily basis in the dataset. The dataset, which is publicly available for research [6], provides reliable and valuable data for comparing the forecasting accuracy among various regression and classification methods. The dataset was incorporated using two database resources. The first database records the fire occurrences in this region that were detected by the inspectors. The second dataset includes several meteorological variables that were collected by the metrological station at the center of Montesinho Natural Park. To improve prediction accuracy, all the variables provided by the second dataset were used. Table 1 describes all the variables from the dataset employed in the experiments. For each forest fire, several attributes, such as the time, date, spatial location, four components of the FWI system, the total burned area and other meteorological information, were registered on a daily basis. In this dataset, spatial locations within a 9×9 grid were specified, and the X - and Y -axes indicate one of the 81 sub-areas obtained from the division of the study area. Temporal variables were also measured as the months of the year and days of the week since the average weather conditions can vary substantially among the 12 months and the days of the week can involve a variety of human activities that have different impacts on the occurrence of forest fires. Also included were the four components of the FWI system that are affected directly by weather conditions, namely the fine fuel moisture code (FFMC), Duff moisture code (DMC), drought code (DC) and initial spread index (ISI). Regarding the meteorological

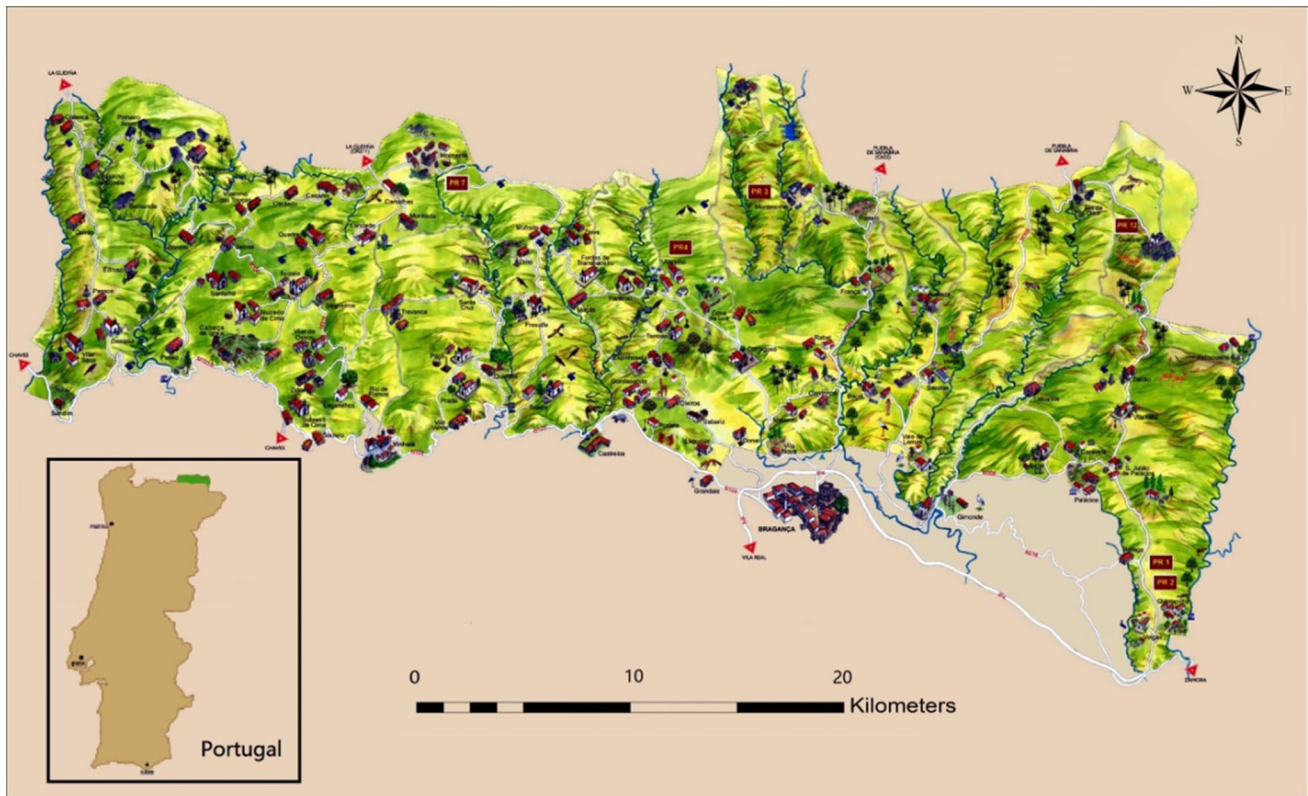


Fig. 1 Map of the Montesinho Natural Park, Portugal

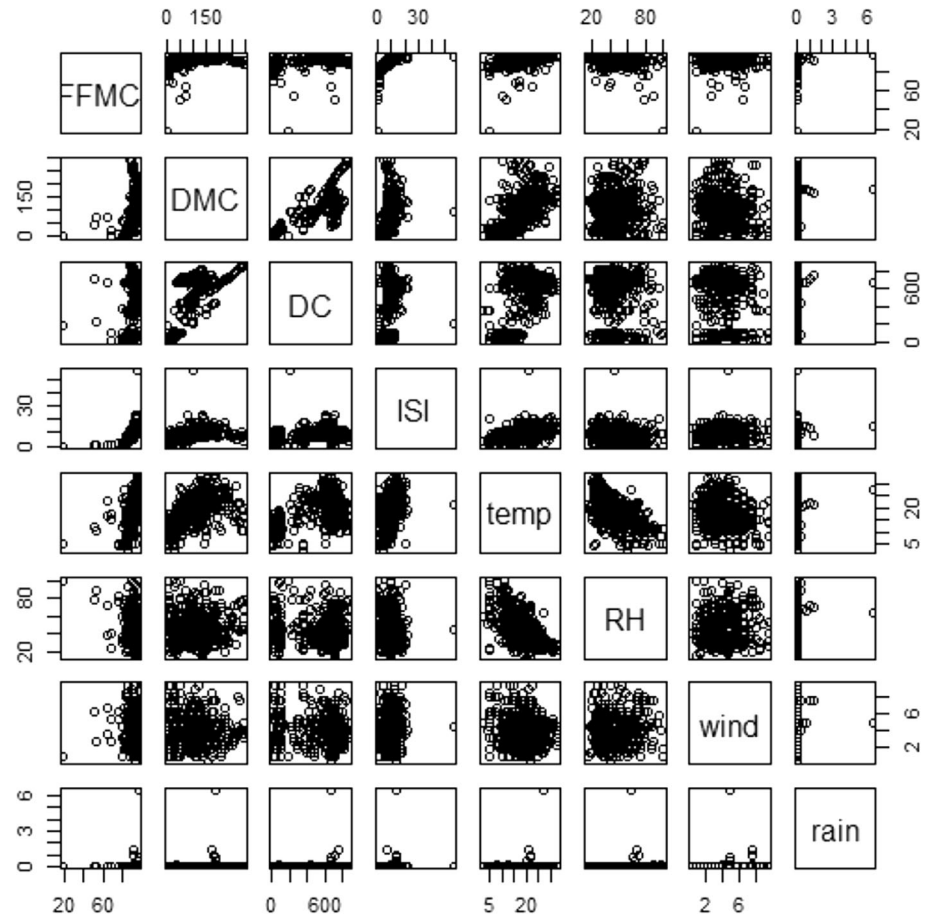
Table 1 Description of input variables and the prediction variable (burned area)

Variables	Description
X	X-axis spatial coordinate (from 1 to 9)
Y	Y-axis spatial coordinate (from 1 to 9)
Month	Month of the year (from “January” to “December”)
Day	Day of the week (from “Monday” to “Sunday”)
FFMC	FFMC code from the FWI system (from 18.7 to 96.20)
DMC	DMC code from the FWI system (from 1.1 to 291.3)
DC	DC code from the FWI system (from 7.9 to 860.6)
ISI	ISI code from the FWI system (from 0 to 56.10)
Temp	Temperature in degrees Celsius (from 2.2 to 33.30)
RH	Relative humidity in percentage (from 15.0 to 100)
Wind	Wind speed in km/h (from 0.40 to 9.40)
Rain	Outside rain in mm/m ² (from 0.0 to 6.40)
Area	Total burned area of the forest (in ha) (from 0.00 to 1090.84)

attributes, four attributes used by the FWI system were selected in the dataset: the temperature, relative humidity, wind speed and precipitation. The values of the first three attributes denote instant records that were obtained by the station’s sensors when the fire was detected. The value of the precipitation was measured as the accumulated precipitation within the previous 30 min. The area variable denotes the burned area of the forest (in ha). Figure 2 plots the correlation among the input variables used in the

analysis, where the DMC is positively and linearly correlated with the DC and the temperature is negatively and linearly correlated with the relative humidity (in percentage format). The scatter plots suggest the high reliability of the data of the input variables that characterize the weather conditions of the FWI system and the meteorological attributes.

Fig. 2 Scatter plot showing the forest fire data of the UCI (Montesinho Natural Park)



5 Results

5.1 Experiment one: regression models predicting the burned area

All the experiments reported in the paper were implemented using H2O,¹ the world's leading open-source DL platform, which facilitates the use of various data mining techniques in regression and classification tasks. To investigate the impact of the input variables, twelve distinct attribute (except “area”) selection configurations, as listed in Table 1, were tested for each data mining algorithm. Several regression models were tested to produce predictions, including the methods of default random forests (RFd), tuned random forests (RFt), default gradient boosting machines (RFd), tuned gradient boosting machines (GBMt), default generalized linear models (GLMd), tuned generalized linear models (GLMt) and DL. Tenfold cross-validation was applied to each model in the experiment, in which the dataset was randomly divided into k subsets, and each model was trained and tested 10 times.

¹ <http://www.h2o.ai/>.

To compare the overall performance of the regression models, the root-mean-square error (RMSE) and the mean absolute error (MAE) were computed to measure how close the forecasts or predictions were to the eventual outcomes. The RMSE can be computed as

$$\text{RMSE} = \sqrt{\sum_{i=1}^n (X_{\text{obs},i} - X_{\text{model},i})^2}$$

where X_{obs} is the observed burned area, and X_{model} is the predicted value of the burned area at time/place i .

The MAE is given by

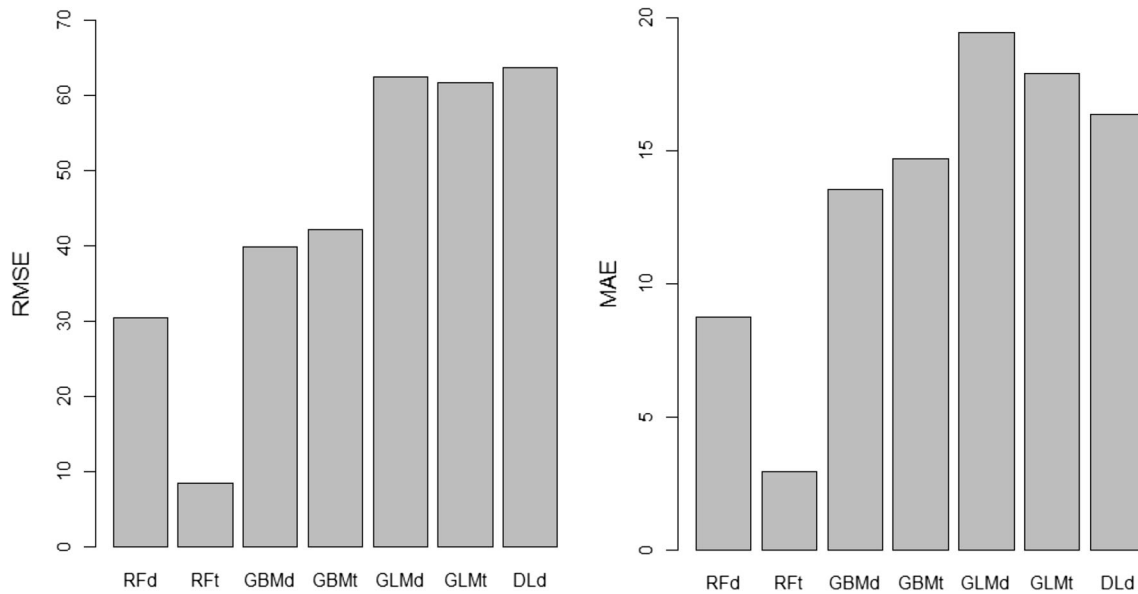
$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - t_i| = \frac{1}{n} \sum_{i=1}^n |e_i|$$

where y_i is the predicted values of the burned area at time/place i (output of the generated model that is evaluated on the training set), and t_i is the corresponding target value of the burned area.

In both metrics, lower values correspond to better predictive performance. RMSE is commonly used and offers an excellent general purpose error metric for predictions. However, relative to the similar MAE, RMSE is more sensitive to large errors. The RMSE and MAE of all the

Table 2 Comparison of the RMSE and the MAE among all the regression models

	RFd	RFt	GBMd	GBMt	GLMd	GLMt	DL
RMSE	30.5031	8.3708	39.8354	42.0884	62.4157	61.7381	63.6423
MAE	8.7392	2.9343	13.5360	14.6881	19.4700	17.9117	16.3923

**Fig. 3** Histograms for the RMSE and MAE for the various models

methods are reported in Table 2, and the histograms for the RMSE and the MAE for the respective model are also provided (Fig. 3).

Under the RMSE and MAE criteria, given the same attribute selection, the RFt² (RMSE = 8.3708 and MAE = 2.9343), as an ensemble learning technique, outperformed other regressions and was likely to make the best prediction of the burned area relative to other traditional GLM models. In addition, under the same criteria, the RMSE and MAE of the RF models (RFt and RFd) were lower than those of models used in the previous study by the UCI machine learning laboratory [6, 9]. The lowest MAE value of the SVM model in the previous study obtained using the same data was 12.71, and the lowest RMSE of the benchmark model was 63.7. The RFt model based on twelve distinct attribute selection setups demonstrated a better predictive ability in forecasting the burned area due to its distinct features. In the RFt model, each randomized tree is built from a sample drawn with replacement from the training set based on the given parameters. This experiment verified the variance decreases and the overall better prediction model that was

generated due to the averaging of the independent and tuned trees.

In the RF model, each tree is constructed from a sample drawn with replacement. During the generation of the tree, the chosen split chosen is the best among a random subgroup of the features (Fig. 4). Because of the randomness, the variance decreases due to the averaging and typically outweighs the increase in bias, which generally leads to improved predictive accuracy [37–39].

5.2 Experiment two: classification models for predicting large-scale fires

Firefighting departments must make reliable decisions regarding resource allocation for fire prevention and suppression if the types of fires with meteorological data can be accurately and automatically classified. In addition, the automatic and accurate classification are expected to be of great benefit to the prevention of the spread of fires and the mitigation of damages to the environment, properties, human lives and livestock. In this regard, large-scale forest fires were arbitrarily recoded as “1” (burned area of > 5 ha) within the dataset. The instances of large-scale forest fires account for almost 30% of the total instances of forest fires in the data used in the study. Then, several machine learning methods, including DTs, RFs, SVMs, NNs, EGB and DL, were implemented with tenfold cross-validation

² The best RFt model setting was “max_depth = 40, ntrees = 200, sample_rate = 0.9, mtries = 4, col_sample_rate_per_tree = 0.9, and score_tree_interval = 10”.

Random Forest

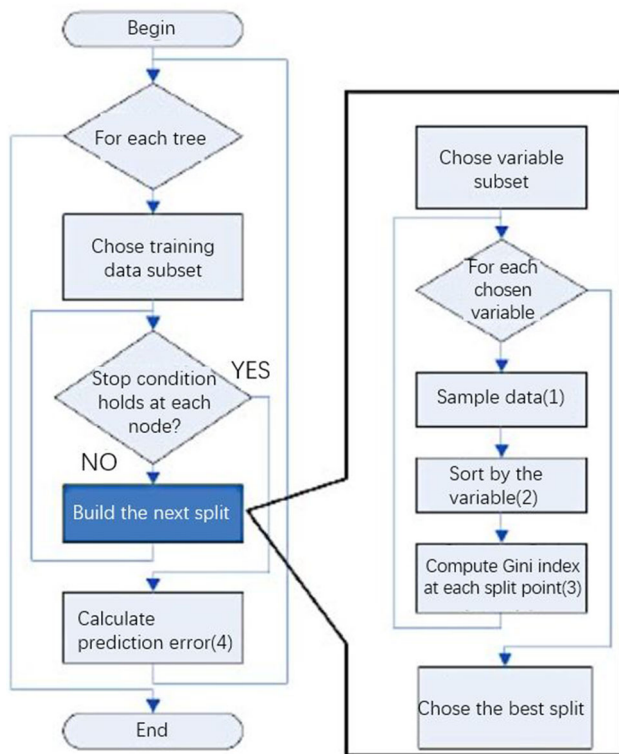


Fig. 4 Flowchart of data processing using the RF method

on each classifier to measure the model's prediction accuracy. For benchmarking purposes, logit regression (LR) was added for comparison with other models. Table 3 lists the values of the correct prediction rate for all the classification models.

Table 3 and Fig. 5 show the correct prediction rate of the seven classification models. Given the same attribute selection, the LR model tended to produce the least accurate prediction of large-scale forest fire occurrences (only 62.5%). Notably, the RF with the best predictive performance in forecasting the burned fire area performed poorly in this experiment. In contrast, the correct prediction rates of the DT, SVM, EGB and DL exceeded 70% with the overall prediction accuracy of the EGB reaching 72.3%, thus enabling the model of EGB³ to outperform the other models in terms of its prediction accuracy for large-scale fires. These results indicate that the EGB method has great potential for the accurate prediction of large-scale forest fire occurrences in other regions around the world. In this experiment, EGB converted weak learners into strong learners by giving more weight to misclassified cases in

³ The EGB settings were as follows: `xgbGrid <- expand.grid(nrounds = c(1, 10), max_depth = c(1, 4), eta = c(.1, .4), gamma = 0, colsample_bytree = .7, min_child_weight = 1, and subsample = c(.8, 1))`.

Table 3 Comparison of prediction accuracy among all the classification models (%)

LR	DT	RF	SVM	NN	EGB	DL
0.625	0.704	0.669	0.709	0.688	0.723	0.712

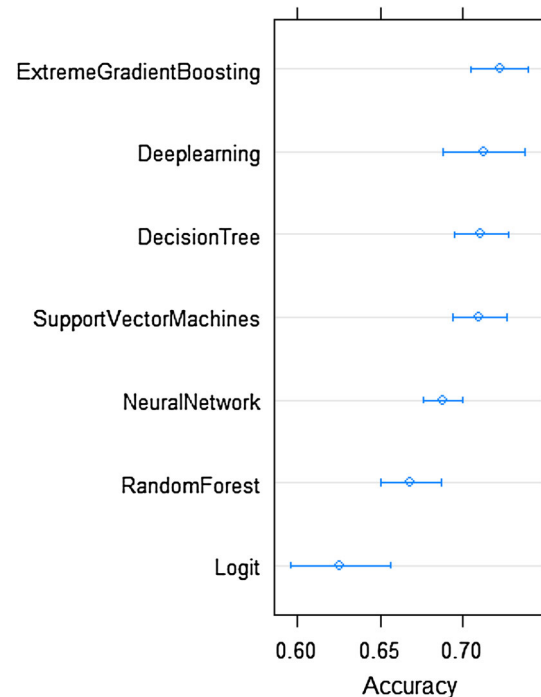


Fig. 5 Prediction accuracy of the seven models

earlier rounds. The method also generated weighted versions of the data. The predictions were combined through a weighted majority vote in the classification process to increase accuracy.

The more accurate predictions generated using EGB result from the data processing that implements a more regularized model formalization to mitigate over-fitting and yield better performance. For each node, the method lists all the features; for each feature, the method ranks the instances according to the feature's importance (Fig. 6). Then, the results are scanned to determine the best split and uses the best split choice along all the features to optimize the sparse data and to approximate a better tree learning solution [40].

6 Conclusions

Forest fires not only threaten the environment but also cause significant damage to property and human lives. In the past decades, substantial effort in the academic

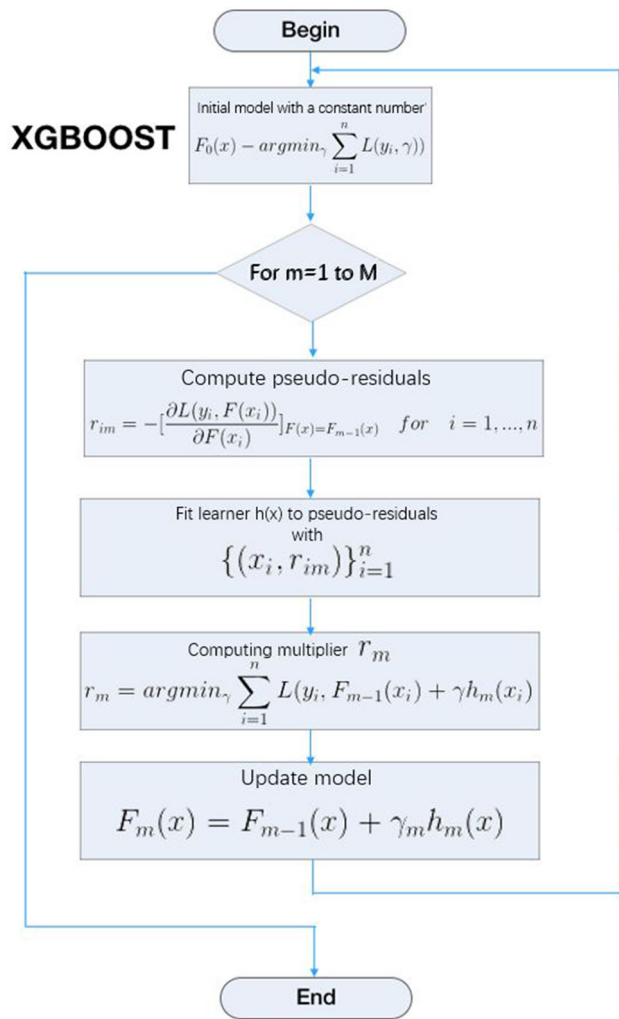


Fig. 6 Flowchart of data processing in the EGB model

literature has been devoted to the development of machine learning methods that assist firefighting management and decision making to conduct reliable and accurate predictions. The main objective of the present study is to present an application of recently emerged ensemble learning methods to the design of automatic and reliable prediction techniques of forest fires by comparing EL methods with other data mining techniques that are commonly used in forest fire predictions. To this end, we used forest fire data from Montesinho Natural Park, Portugal that form a widely used benchmark for empirical evaluation of new and existing learning algorithms to calculate and compare the performance of different prediction models. The results of various regression models and classification models were reported according to the potential practical demands. Relative to data mining techniques, the Rf approach performed better than the other regression models did with regard to the prediction accuracy of the burned area in terms of a comparatively small RMSE and MAE obtained

from the experimental results. In the method, each randomized tree was built from a sample drawn with replacement from the training set according to the given parameters. The experiment verified the decrease in variance and the general improvement in the prediction accuracy due to the averaging of the independent and tuned trees. In addition, the EGB method outperformed the other classification models in terms of the predictive accuracy of large-scale fire occurrences by giving more weight to misclassified cases in earlier rounds and converting weak learners to strong learners. In this model, the predictions were combined through a weighted majority vote in the classification process to produce more accurate results.

Predicting the occurrences of forest fires and burn areas is a challenging task. More accurate prediction techniques would be of particular significance in strong fire seasons when simultaneous fires may occur at various locations. The findings show that ensemble learning methods not only have great potential for broader applications in forest fire automatic precaution and prevention systems but also can provide important techniques for forest firefight decision making in terms of fire resource allocation and strategies to build proactive responses (including fire towers, inspection stations and fire patrols) and can ultimately improve the efficiency of forest fire management worldwide.

Acknowledgements This research was financially funded by the Ministry of Education in China (MOE)'s Project of Humanities and Social Sciences (Project No. 15YJCZH128).

Author contributions Ying Xie and Minggang Peng conceived and designed the experiment and collected and analyzed the data. Minggang Peng and Ying Xie wrote the original draft.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

1. Food and Agriculture Organization of the United Nations (2015) Global forest resources assessment. www.fao.org/3/a-i4808e.pdf. Accessed 1 Feb 2017
2. Brushlinsky NN, Hall JR, Sokolov SV, Wagner P (2016) World fire statistics. www.ctif.org/ctif/world-fire-statistics. Accessed 1 Feb 2017
3. Stephens SL (2005) Forest fire causes and extent on United States Forest Service lands. *Int J Wildland Fire* 14:213–222. <https://doi.org/10.1071/WF04006>
4. Guo F, Innes JL, Wang G, Ma X, Sun L, Hu H, Su Z (2015) Historic distribution and driving factors of human-caused fires in the Chinese boreal forest between 1972 and 2005. *J Plant Ecol* 8:480–490. <https://doi.org/10.1093/jpe/rtu041>
5. Pyne SJ, Andrews PL, Laven RD (1996) *Introduction to wildland fire*, 2nd edn. Wiley, London
6. Cortez P, Morais AdJR (2007) A data mining approach to predict forest fires using meteorological data. In: Neves JM, Santos MF,

- Machado JM (eds) New trends in artificial intelligence: proceedings of the 13th Portuguese conference on artificial intelligence (EPIA 2007), Guimarães, Portugal, [Lisboa]: APPIA, 2007. ISBN 978-989-95618-0-9, 2007. pp 512–523
7. Arrue BC, Ollero A, de Dios JRM (2000) An intelligent system for false alarm reduction in infrared forest-fire detection. *IEEE Intell Syst* 15:64–73. <https://doi.org/10.1109/5254.846287>
 8. Piñol J, Terradas J, Lloret F (1998) Climate warming, wildfire hazard, and wildfire occurrence in coastal eastern Spain. *Clim Change* 38:345–357. <https://doi.org/10.1023/A:1005316632105>
 9. Castelli M, Vanneschi L, Popović A (2015) Predicting burned areas of forest fires: an artificial intelligence approach. *Fire Ecol* 11:106–118. <https://doi.org/10.4996/fireecology.1101106>
 10. Angayarkkani K, Radhakrishnan N (2009) Efficient forest fire detection system: a spatial data mining and image processing based approach. *Int J Comput Sci Netw Sec* 9:100–107
 11. Pourtaghi ZS, Pourghasemi HR, Aretano R, Semeraro T (2016) Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques. *Ecol Ind* 64:72–84. <https://doi.org/10.1016/j.ecolind.2015.12.030>
 12. Hr Pourghasemi, Beheshtirad M, Pradhan B (2016) A comparative assessment of prediction capabilities of modified analytical hierarchy process (M-AHP) and Mamdani fuzzy logic models using Netcad-GIS for forest fire susceptibility mapping. *Geomat Nat Hazards Risk* 7:861–885. <https://doi.org/10.1080/19475705.2014.984247>
 13. Saoudi M, Bounceur A, Euler R, Kechadi T (2016) Data mining techniques applied to wireless sensor networks for early forest fire detection. In: proceedings of the international conference on internet of things and cloud computing. ACM, New York. p 71
 14. Stojanova D, Panov P, Kobler A, Džeroski S, Taškova K (2006) Learning to predict forest fires with different data mining techniques. In: Conference on data mining and data warehouses, Ljubljana
 15. Wijayanto AK, Sani O, Kartika ND, Herdiyeni Y (2017) Classification model for forest fire hotspot occurrences prediction using ANFIS algorithm. *IOP Conf Ser Earth Environ Sci* 54:012059. <https://doi.org/10.1088/1755-1315/54/1/012059>
 16. Amatulli G, Rodrigues MJ, Trombetti M, Lovreglio R (2006) Assessing long-term fire risk at local scale by means of decision tree technique. *J Geophys Res Biogeosci*. <https://doi.org/10.1029/2005JG000133>
 17. Vasconcelos M, Silva S, Tomé M, Alvim M, Pereira J (2001) Spatial prediction of fire ignition probabilities: comparing logistic regression and neural networks. *Photogramm Eng Remote Sens* 67:73–81
 18. Vega-Garcia C, Lee BS, Woodard PM, Titus SJ (1996) Applying neural network technology to human-caused wildfire occurrence prediction. *AI Appl* 10:9–18
 19. Mazzoni D, Tong L, Diner D, Li Q, Logan J (2005) Using MISR and MODIS data for detection and analysis of smoke plume injection heights over North America during summer 2004. *AGU Fall Meet Abstr* 1:853
 20. Massada AB, Syphard AD, Stewart SI, Radeloff VC (2012) Wildfire ignition-distribution modelling: a comparative study in the Huron-Manistee National Forest, Michigan, USA. *Int J Wildland Fire* 22:174–183. <https://doi.org/10.1071/WF11178>
 21. Zhou ZH (2015) Ensemble learning. In: Li SZ, Jain A (eds) *Encyclopedia biometrics*. Springer, New York, pp 411–416
 22. Johnson NE, Ianiuk O, Cazap D, Liu L, Starobin D, Dobler G, Ghandehari M (2017) Patterns of waste generation: a gradient boosting model for short-term waste prediction in New York City. *Waste Manag* 62:3–11. <https://doi.org/10.1016/j.wasman.2017.01.037>
 23. Chuvieco E, Aguado I, Jurdao S, Pettinari ML, Yebra M, Salas J, Hantson S, Riva J, Ibarra P, Rodrigues M, Echeverría M, Azqueta D, Román M, Bastarrika A, Martínez S, Recondo C, Zapico E, Martínez-Vega J (2012) Integrating geospatial information into fire risk assessment. *Int J Wildland Fire* 23:606–619. <https://doi.org/10.1071/WF12052>
 24. Loepfe L, Martínez-Vilalta J, Piñol J (2011) An integrative model of human-influenced fire regimes and landscape dynamics. *Environ Model Softw* 26:1028–1040. <https://doi.org/10.1016/j.envsoft.2011.02.015>
 25. Rodrigues M, de la Riva J (2014) An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environ Model Softw* 57:192–201. <https://doi.org/10.1016/j.envsoft.2014.03.003>
 26. Hsu CW, Chang CC, Lin CJ (2003) A practical guide to support vector classification. Technical report, department of computer science and information engineering, University of National Taiwan, Taipei
 27. Zhang C, Ma Y (2012) *Ensemble machine learning: methods and applications*. Springer, Berlin
 28. Ishwaran H, Kogalur UB (2010) Consistency of random survival forests. *Stat Probab Lett* 80:1056–1064. <https://doi.org/10.1016/j.spl.2010.02.020>
 29. Breiman L (2010) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
 30. Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. *Ecology* 88:2783–2792
 31. Ghafouri-Kesbi F, Rahimi-Mianji G, Honarvar M, Nejati-Javaremi A (2017) Predictive ability of random forests, boosting, support vector machines and genomic best linear unbiased prediction in different scenarios of genomic evaluation. *Anim Prod Sci* 57:229–236. <https://doi.org/10.1071/AN15538>
 32. Archibald S, Roy DP, Wilgen BWV, Scholes RJ (2009) What limits fire? An examination of drivers of burnt area in Southern Africa. *Glob Change Biol* 15:613–630. <https://doi.org/10.1111/j.1365-2486.2008.01754.x>
 33. Wu Z, He HS, Yang J, Liu Z, Liang Y (2014) Relative effects of climatic and local factors on fire occurrence in boreal forest landscapes of Northeastern China. *Sci Total Environ* 493:472–480. <https://doi.org/10.1016/j.scitotenv.2014.06.011>
 34. Oliveira S, Oehler F, San-Miguel-Ayán J, Camia A, Pereira JMC (2012) Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *For Ecol Manag* 275:117–129. <https://doi.org/10.1016/j.foreco.2012.03.003>
 35. Li C (2014) A Gentle introduction to gradient boosting. http://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/slides/gradient_boosting.pdf. Accessed 1 Feb 2017
 36. Chen TQ, He T (2015) Xgboost: extreme gradient boosting. R package version 0.4-2. <http://cran.fhrc.org/web/packages/xgboost/vignettes/xgboost.pdf>. Accessed 1 Feb 2017
 37. Liaw A, Wiener M (2002) Classification and regression by random forest. *R News* 2:18–22
 38. Louppe G, Geurts P (2012) Ensembles on random patches. In: Flach PA, Bie TD, Cristianini N (eds) *Machine learning and knowledge discovery in databases*. Springer, Berlin, pp 346–361
 39. Arlot S, Genuer R (2014) Analysis of purely random forests bias. arXiv preprint arXiv 1407:3939
 40. Chen TQ, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, pp 785–794