A MACHINE LEARNING METHODOLOGY FOR NEXT DAY WILDFIRE PREDICTION

Stella Girtsou [1], Alexis Apostolakis[1], Giorgos Giannopoulos[2], Charalampos Kontoes[1]

[1]National Observatory of Athens, Athens, Greece {sgirtsou, alex.apostolakis, kontoes}@noa.gr
[2]ATHENA Research Center, Marousi, Greece giann@athenarc.gr

IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9554301

ABSTRACT

In this paper, we handle the problem of next day wildfire prediction via the use of machine learning. In contrast to most works in the relevant literature, we set the problem to its realistic basis, with respect to its large scale, the extreme imbalance in the data distribution, the required high spatial granularity of the predictions and the consideration of the strong spatial correlations inherent in the data. We implement a machine learning workflow that exploits Tree Ensemble and Neural Network algorithms, upon which an extensive hyperparameter search procedure is performed, via cross-validation, in order to select a set of effective models that are expected to generalize well on new data. Our experiments on the whole Greek territory demonstrate the effectiveness of the proposed methodology, rendering it directly applicable to real-world scenarios. Finally, several insights towards further improving the effectiveness of current models are discussed.

Index Terms— Wildfire prediction, next day prediction, machine learning, Tree Ensembles, Neural Networks

1. INTRODUCTION

Wildfires are events with catastrophic impacts on environment, economy and, on several occasions, on people's lives. Especially in the Mediterranean territory, wildfires have been causing the loss of human lives, huge environmental disaster, as well as subsequent socio-economic costs in parceling, compensations and medical expenses. Developing methods for the timely prediction of wildfires events or, at least, for grading the risk of fire events per territory has been gaining increasing interest by the research community. Lately, machine learning (ML) has become the de-facto methodology applied in the task. However, most approaches in the literature (see [1] for an extensive review) either solve a simplified version of the problem or present methodological shortcomings, limiting thus their applicability/generalizability in real-world settings.

An important omission in several works ([2]–[6]) is that they ignore the extremely imbalanced data distribution. Considering, e.g., the Greek territory, one of the most prone countries to wildfires, the ratio of fire to non-fire areas is ~1:100,000. Approaches adopting a balanced dataset setting, although often demonstrating impressive prediction results, are essentially detached from a real-world setting.

An additional shortcoming of existing works consists in the poor exploration of the algorithms' hyperparameter spaces, which, if properly configured, can significantly change the algorithms' effectiveness. However, selecting a proper hyperparameter configuration is a highly data-driven process, and might lead to completely different configurations for different datasets. Most works [2]–[4], [6]–[8] select either the default configuration of each algorithm or the best performing one after a limited trial-and-error search, which, however, might be sub-optimal in the context of the task and the underlying data.

Another significant shortcoming of several state of the art works ([4], [8], [9]) regards the spatio-temporal correlation of the instances (areas of a territory). Ignoring this fact by, e.g., performing shuffling of the data before learning and assessing a machine learning algorithm, leads to misleading effectiveness results, since it allows almost identical instances to be shared between training and test sets, essentially performing data leakage.

In this paper, we handle the problem of next day wildfire prediction on areas (grid cells) of a territory, considering a particularly fine-grained partition grid, with cells 500m wide. To this end, we select a set of state of the art algorithms (Random Forest, XGBoost, LogitBoost and Neural Networks) and tune them by searching their hyperparameter spaces via cross-validation on a historical training dataset. The best performing models are selected via this process with respect to different evaluation measures (AUC, F-score of fire class, harmonic mean of Recalls from both classes). These models are eventually assessed on a separate real-world (extremely imbalanced) test set. The reported results demonstrate the applicability of the proposed method in real-world deployment settings, as well as provide us with insights towards further improving the effectiveness of the proposed methods. Building on and extending our previous work in [10], the main contributions of the paper are as follows:

We formalize the problem on its most realistic basis, focusing on next day prediction, providing predictions on a highly granular scale and assessing the proposed methods on a real-world (extremely imbalanced) test set.

We present a sound methodology for tuning, comparing, selecting and deploying ML models on the task, that ensures avoiding the pitfalls of overfitting and data leakage.

We identify and discuss the specificities of the task (imbalance, spatio-temporal correlations) proposing directions towards further improving the proposed methods.

2. MATERIALS AND METHODS

Next, we first provide the problem definition and then outline the proposed method, describing the training features, the ML algorithms, and the proposed machine learning workflow.

2.1. Problem definition

In this work, we handle the problem of next day fire prediction. Our dataset comprises a geographic grid of high granularity (each cell being 500m wide) covering the whole territory of interest (in our case, the whole Greek territory). Each instance corresponds to the daily snapshot of each grid cell, and is represented by a set of characteristics (features) that are extracted for the specific area, for a specific day d_k. Given a historical dataset annotated (labeled) with the existence or absence of fire, for each grid cell, for each day, each available historical instance carries a binary label l_k for the day d_k, denoting the existence (label:fire) or absence of fire (label:no-fire). It is important to point out that all the features of the instance d_k, are *available from the previous day* d_{k-1} because they either (1) are invariant in time, (2) have a slow variation (3) or can be represented with satisfying accuracy by next day predictions. Thus, our problem is formulated as a binary classification task; the goal is to learn, using historical data, a decision function $f(x_k;\theta)$ that, given a new instance x_k accurately predicts label l_k.

2.2. Training features and ML algorithms

A fire inventory for the years 2010-2019 was constructed, obtaining data from multiple sources including FIREHUB, NOA, EFFIS and NASA as explained in [10]. This includes daily snapshots of cells (areas) of the aforementioned geographic grid (Greek territory), from March until October, limiting the study region to forest areas. Upon this dataset, 15 fire driving factors were extracted to serve as features within the deployed ML algorithms. These features are summarized below, while a more detailed description is provided in [10].

NDVI/EVI. These comprise two quantitative features regarding vegetation of an area. The EVI feature is added in the current work, compared to [10].

Temperature. These comprise three quantitative weather features (max_temp, min_temp, mean_temp).

Wind components. These comprise two quantitative (max_wind_speed, max_wind_speed_dominant_direction) and two categorical (wind_direction, dominant_direction) weather features. In the current work, compared to [10], we represent the above categorical features via one-hot encoding.

Precipitation. This comprises 1 quantitative weather feature.

Land use/cover. This comprises one categorical land cover feature, similarly re-represented via one-hot encoding compared to [10].

DEM. These comprise four quantitative topographic features (DEM, aspect, slope, curvature).

Subsequently, several data cleaning operations were applied on the datasets, in particular: (a) Each missing value was replaced with the mean of the N closest cells with non-missing values; (b) Outliers were trimmed in order to avoid our models to be impacted from extreme values; (c) Standardization and normalization were applied to ensure the optimal execution for some of the adopted algorithms (Neural Networks); (d) Categorical features were encoded in one-hot encoding. Thus, the dimensionality of our feature vectors increased to 43.

We considered state of the art algorithms with consistently high performance in similar classification tasks [1]. In particular, we adopted Tree Ensemble algorithms either based on bagging (Random Forest - RF) or on boosting (XGBoost - XGB, LogitBoost - LGB) and shallow Neural Networks – NN architectures of maximum 3 hidden layers. A large space of hyperparameters for each of these four algorithms was explored via the cross-validation process described next, which is omitted due to lack of space.

2.3. Machine learning workflow

An extensive hyperparameter search for each algorithm is performed via a 10-fold cross-validation [11] scheme in order to identify the best performing models. In each iteration, k-1 folds are used for training, leaving the remaining fold for validation. The best performing models were selected based on the average validation performance on all folds with respect to the measures of *F-score*, *ROC-AUC*, as well as *hybrid measures* (*HMRa* and *HMRb*) that are derived by the adjusted *harmonic mean* of the *Recall* measures for both prediction classes. Our target was to explore the most fitting measure for maximization during cross-validation for our scenario. HMRa and HMRb were defined as follows:

- HMRa = 3/(2/Recal1+1/Recal0)
- HMRb = 6/(5/Recal1+1/Recal0)

where Recall1 and Recall0 correspond to the Recall values for the fire and no-fire classes respectively. The hybrid measures aim to favor the fire class Recall and also preserve a relatively high level of no-fire class Recall.

To account for the strong spatial correlations in the data, which could lead to data leakage and model overfitting, a strict rule was followed in the 10-fold splitting process: cells of a specific day were not allowed to be distributed in more than one fold. This rule effectively prevented neighboring cells from the same day and *the same fire event*, to be included in both the training and the validation folds during cross-validation. Omitting this rule would probably produce validation partitions easier to predict but it would also compromise the model generalization capability.

Further, given the massive amount of training data (~550M instances), and the heavyweight processing required during the hyperparameter search in a cross-validation scheme, we perform undersampling on the initial training dataset, and produce a balanced training set of ~25K

instances. We note that this dataset is only utilized for model tuning and selection, via the procedure described above; the selected models are eventually assessed on a hold-out test set, which maintains the real-world, extremely imbalanced distribution of classes (~1:100K ratio of fire/no-fire).

Another important consideration is to be able to balance the importance of fire and no-fire misclassifications on the learned models. To this end, out of the set of the searched hyperparameters, the "class weights" parameter has particular importance. Note that, the class no-fire on a cell of our dataset is to an extent misleading, in the sense that several of those cells, with close characteristics to actual fire cells, also have a high risk of fire; the fact that a fire did not occur could be attributed to factors that are infeasible to capture by any feature set, irrespective of how elaborate they might be (e.g. a random driver tossing a cigarette at a random location). Thus, it is more important for a model to predict such cells as "fire" rather than "no-fire", even though labeled otherwise in the training dataset, with a consequent cost on the Recall0 measure. Via the "class weights" hyperparameter, we are able to force the training algorithm to explore different balancing ratios on the importance of fire and no-fire misclassifications, further facilitating different model tuning for the different considered measures during the cross-validation process.

The outcome of the above process provides four trained models (one for each measure) for each considered algorithm. Next, we present their performance on the hold out test set.

3. RESULTS

In this section, we report the evaluation results of the assessed models on the hold-out test set, comprising 386 fire and 11,687,126 no-fire cells for August 2019. We note that in the training set that was used in the cross-validation process we have only considered data from years 2010-2018. The following table presents **Recall values** for no-fire and fire classes, achieved by each of the four algorithms (NN, RF, XGB, LGB) on both the validation (averaged over the 10 folds) and the test set, for models tuned on the four different measures during cross-validation (Tuned by AUC/F-score/ HMRa/HMRb).

We observe that the Recall scores for class fire are maintained or even increased near or above 0.9 for the test set in comparison to the validation set average, which implies a robust workflow and models with high generalizability. It is noticeable that the employment of the HMR measures boosted the fire class detection and led to models with near-excellent recall on fires (96% and 97% Recall for RFs and NNs respectively). On the contrary, the scores of no-fire class in the test set are much lower than the corresponding average scores of the training set. This was expected due to the huge imbalance of the real-word test set and also because the training was performed on a modified distribution since that training set contained only a small subset of the actual no-fire instances. These findings support our claim that reporting test results on a balanced dataset can be misleading; the proposed models need to be assessed on test sets that maintain the real-world distribution, so that reliable conclusions can be made.

Table 1: Best Models recall for no fire (0) and fire (1) class

Tuned by:	AUC		F-score		HMRa		HMRb	
Class / Alg/ Dataset	No Fire	Fire	No Fire	Fire	No Fire	Fire	No Fire	Fire
NN/Valid.	0,79	0,78	0,66	0,90	0,72	0,86	0,65	0,91
NN/Test	0,42	0,87	0,25	0,93	0,39	0,87	0,22	0,97
RF/Valid.	0,67	0,86	0,61	0,91	0,58	0,93	0,52	0,95
RF/Test	0,36	0,92	0,28	0,94	0,26	0,95	0,23	0,96
XG/Valid.	0,60	0,88	0,55	0,93	0,55	0,93	0,46	0,97
XG/Test	0,37	0,89	0,36	0,92	0,36	0,90	0,39	0,91
LB/Valid.	0,98	0,54	0,98	0,54	0,98	0,54	0,98	0,54
LB/Test	0,51	0,71	0,51	0,71	0,51	0,71	0,51	0,71

An interesting performance is observed for the best LGB model. The Recall of the no-fire class is quite high for that model even though the Recall of fire is low. The different behavior of this algorithm makes it suitable for being combined as an "ensemble" with one of the other models, in order to improve the final scoring.

Further, it is evident that the assessed algorithms and measures for tuning during cross-validation provide a degree of flexibility as to which models to choose for deployment: if missing as less fires as possible is the only criterion, in expense of predicting the majority of a territory as fire, then a NN model tuned on HMRb is optimal; if limiting the territory that is predicted as fire is also important, even if this means missing some actual fires, a NN tuned on AUC or even LGB are preferable; if a middle ground is preferred, then several XGB or RF configurations can be selected.

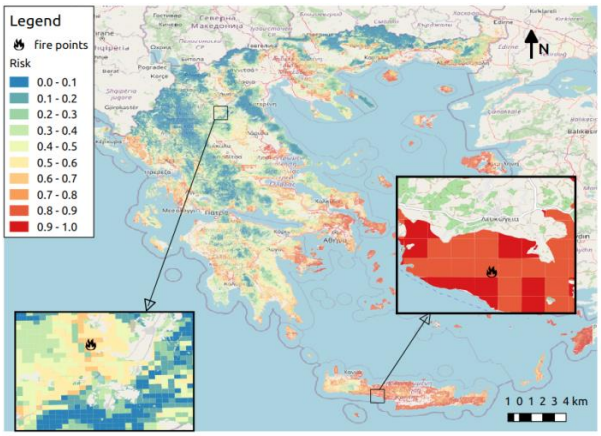

Figure 1: Risk map for 18-08-2019

Finally, Figure 1 presents an example of visualization of the prediction results of the best NN model for one day of the test set, the 18th of August 2019. The color map values correspond to the two node output layer values of the NN

(0.0-1.0) that represent the model's probabilities to select each class. Making the assumption that this probability is equal to the fire risk we can regard this map as a fire risk map for the specific day. The two fire events that were actually recorded were both on cells where the risk score was above 0.6. In the case of the fire in the south (right zoomed panel) the whole area had high risk scores (>0.8) while in the case of the fire in the north of the mainland (left zoomed panel) even though many cells around the fire had low risk scores, the fire event started and evolved within the few cells that had risk scoring >0.6. The latter case is also an indication of the practicality of the high resolution (500m cell) of the map.

4. DISCUSSION AND FUTURE WORK

In the previous sections, we presented a machine learning methodology for efficient, effective and sound learning of models for next day fire prediction. Further, we demonstrated its effectiveness on a real-world setting covering the Greek territory. Nevertheless, our experiments, as well as the ongoing exploration and analysis of the data at hand indicate several directions for improving the underlying models and the proposed methodology, as discussed next.

Although the proposed cross-validation scheme ensures that no data leakage takes place, as well as the reliability of the reported measures on the test set, it is still sub-optimal for our task. The reason is that the best models are selected on balanced validation sets, thus on a modified distribution compared to the real-world. Our ongoing work further improves this process by keeping the balanced distribution only for the heavyweight part of cross validation, i.e. model training, while it considers imbalanced validation partitions for model validation-tuning. This way, the optimal models are selected based on their effectiveness on validation sets that actually have the same distribution with the real-world test set. We expect that the above configuration will lead to increase of no-fire Recall, without sacrificing fire Recall.

Another direction concerns the further examination of the inherent spatial and temporal correlations of the data. Deep learning algorithms in the family of Convolutional NN and Recurrent NN (especially LSTM) [12] have shown great capability in discovering spatial and temporal correlations respectively, thus we intend to explore such techniques in our future work to better exploit these correlations.

Finally, it is evident that additional training features regarding, e.g., the relative location and fire history of an area, infrastructures, road network, soil moisture, land surface temperature, demographics and features that imply anthropogenic activities could potentially improve the effectiveness of the models. Further, an extended feature space could be better exploited by deeper NN architectures, as well as specialized architectures, such as Siamese Networks for outlier detection (since the extreme imbalance in the data resembles an outlier detection setting).

ACKNOWLEDGEMENTS

This paper has been supported by the Action titled "National Network on Climate Change and its Impacts - Climpact" which is implemented under the sub-project 3 of the project ""Infrastructure of national research networks in the fields of Precision Medicine, Quantum Technology and Climate Change"", funded by the Public Investment Program of Greece, General Secretary of Research and Technology/Ministry of Development and Investments

REFERENCES

[1] P. Jain, S. C. P. Coogan, S. G. Subramanian, M. Crowley, S. Taylor, and M. D. Flannigan, "A review of machine learning applications in wildfire science and management," 2020.
[2] M. Tonini, M. D'andrea, G. Biondi, S. D. Esposti, A. Trucchia, and P. Fiorucci, "A machine learning-based approach for wildfire susceptibility mapping. The case study of the liguria region in italy," Geosci., vol. 10, no. 3, pp. 1–18, 2020.
[3] L. Gigovi´cgigovi´c, H. R. Pourghasemi, S. Drobnjak, and S. Bai, "Testing a New Ensemble Model Based on SVM and Random Forest in Forest Fire Susceptibility Assessment and Its Mapping in Serbia's Tara National Park," 2019.
[4] M. Rodrigues and J. De la Riva, "An insight into machine-learning algorithms to model human-caused wildfire occurrence," Environ. Model. Softw., vol. 57, pp. 192–201, Jul. 2014.
[5] A. Alonso-Betanzos et al., "An intelligent system for forest fire risk prediction and fire fighting management in Galicia," Expert Syst. Appl., vol. 25, no. 4, pp. 545–554, Nov. 2003.
[6] M. S. Tehrany, S. Jones, F. Shabani, F. Martínez-Álvarez, and D. Tien Bui, "A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using LogitBoost machine learning classifier and multi-source geospatial data," Theor. Appl. Climatol., vol. 137, no. 1–2, pp. 637–653, Jul. 2019.
[7] A. A. Bar Massada, A. D. Syphard B, S. I. Stewart C, and V. C. Radeloff A, "Wildfire ignition-distribution modelling: a comparative study in the Huron-Manistee National Forest, Michigan, USA.
[8] G. Zhang, M. Wang, and K. Liu, "Forest Fire Susceptibility Modeling Using a Convolutional Neural Network for Yunnan Province of China," Int. J. Disaster Risk Sci., vol. 10, no. 3, pp. 386–403, Sep. 2019.
[9] M. Bisquert, E. Caselles, J. M. Snchez, and V. Caselles, "Application of artificial neural networks and logistic regression to the prediction of forest fire danger in Galicia using MODIS data," Int. J. Wildl. Fire, vol. 21, no. 8, pp. 1025–1029, 2012.
[10] A. Apostolakis, S. Girtsou, C. Kontoes, I. Papoutsis, and M. Tsoutsos, "Implementation of a Random Forest classifier to examine wildfire predictive modelling in Greece using diachronically collected fire occurrence and fire mapping data (forthcoming)," in 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II, J. Lokoč, T. Skopal, K. Schoeffmann, V. Mezaris, X. Li, S. Vrochidis, and I. Patras, Eds. .
[11] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," in Encyclopedia of Database Systems, L. LIU and M. T. ÖZSU, Eds. Boston, MA: Springer US, 2009, pp. 532–538.
[12] W. Yuankai and T. Huachun, "Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework," 2016.