Automatic machine learning Framework for Forest fire forecasting

Jintao Qu1,2, Xiaohui Cui1,2*

1School of Information Science and Technology, Beijing Forestry University, Beijing, 100083, China
2Engineering Research Center for Forestry-orientend Intelligent Information Processing, National Foreatry and Grassland Administration, Beijing, 100083, China
*Corresponding author’s e-mail: qujintao@bfuer.com

Abstract. Based on the automatic machine learning framework, combined with the characteristics of forest fire meteorological data and the adaptive requirements of forest fire prediction, this paper optimizes the data preprocessing, parameter learning, loss function and other links of auto-sklearn, builds a forest fire risk prediction framework with regional adaptive characteristics. Based on the forest meteorological fire risk data, a forest fire risk prediction model with regional characteristics and self-learning characteristics is constructed to solve the problems of low compatibility of the existing machine learning methods with binary unbalanced forest fire data, improve the accuracy of forest fire prediction and provide decision-making basis for forestry risk management. The comparative analysis results show that the prediction accuracy of the improved framework in different test sets is improved by 13% on average. Compared with the existing machine learning model for forest fire prediction, the prediction accuracy of the framework proposed in this paper is comprehensively better than the existing methods in terms of real forest fire data.

1. Introduction
Forest fires are sudden and destructive. Affected by natural factors and limited by personnel scheduling, it is less likely to be extinguished in a short time. Therefore, accurate prediction and effective prevention is important for forest fire management[1].

Among many factors that may cause forest fires, weather and climate conditions have always been considered as the most important influencing factors. For example, the Canada's forest fire weather index system, which is currently the most developed and widely used in the world. It calculates the forest fire whether index(FWI) based on features such as temperature, humidity and wind speed. The United States, New Zealand and some others have also successfully formed their own fire danger level forecast systems based on the ideas of the FWI.

Although there are many mature forecasting systems in the world, there are differences in climate in different regions. Even in the same forest region, there are some differences in combustibles, topography and other factors in different district. This results in that the parameter setting of the model depends on the assistance of experts during the regionalization application of the prediction model. The complexity of technology limits its wide application at the grass-roots level.

In this paper, the automatic machine learning technology is applied to the prediction of forest fire danger, and its advantages in automatic modeling are fully exploited. Reduces the cost of the model
training and meets the needs of regional adaptation of the forest fire prediction model. In addition, this paper optimizes the existing framework specifically according to the characteristics of unbalanced distribution of the forest fire dataset samples, so as to further improve the regional prediction accuracy of forest fire. The experimental results show that the optimized framework not only performs well in the field of forest fire prediction, but also has great potential in other unbalanced data classification scenarios.

2. Principles of automatic machine learning
The automatic machine learning framework takes Bayesian optimization algorithm as the core method to realize the automation of feature engineering, classifier selection and hyper-parameter adjustment for various datasets[2]. On this basis, auto-sklearn draws lessons from the ideas of experts in selecting models and parameter tuning according to different historical datasets and experience, and then adds meta-learning process. That is, extract the features of the historical training dataset and store the classifier and hyper-parameters with better performance, so that the features of the new dataset can be analyzed before the new learning task and inspire the subsequent optimization process. In this way, the efficiency of model training can be greatly improved. In addition, all models and hyper-parameters in the Bayesian optimization process are retained, and fast integrated learning is carried out in the later stage to reduce the error rate of a single model and improve the overall performance of the model[3].

3. Optimization for classification of binary unbalanced datasets
3.1. Incremental extension of empirical data sets
Just as experts have accumulated experience in the performance and application of machine learning algorithms from previous projects, meta-learning selects the model that may be most suitable for the target dataset by reasoning the performance of algorithms in other datasets. Therefore, accumulating sufficient prior data sets is the key to improving meta-learning performance.

Auto-sklearn has 136 prior data sets, while there are only 8 binary unbalanced prior data sets, and the meta feature coverage of the prior data sets is incomplete[4]. Based on the original framework, the new framework collects 12 additional unbalanced datasets in the OpenML, the high-performance classifier models and parameter tuning results selected by many machine learning experts are marked as super-parameters of these datasets. By introducing more prior data, the meta-learning process can match the target dataset with the prior dataset more accurately, and select more appropriate models and super-parameters, the initial classification accuracy of Bayesian optimization process is improved, which helps the model converge faster and thus improves the classification accuracy of the final binary unbalanced dataset.

3.2. Recognition and matching of binary unbalanced datasets
We extract the category of the target data set from the multidimensional features and perform independent detection. If the data set is a binary data set, the range of objects whose meta-learning completes k nearest neighbor matching will be reduced to a binary prior data set. Thereafter, if it is determined that the data set is unbalanced, the meta-learning data set is further simplified into a binary unbalanced data set to ensure that the target data set can maximize the use of the parameters of the
previous binary unbalanced data set. On the contrary, if the data is a binary balanced data set, only the meta-learning matching object is simplified to a binary prior data set and matched with other meta-features; if the data set is neither binary data nor unbalanced data, you can perform the original frame matching process as usual.

3.3. Optimization of Bayesian optimization process
After Bayesian optimization is completed and new hyper-parameters are obtained, the new parameters need to be introduced into the model for classification. The classification effect of the new model is determined through the loss function, and compared with the loss value. If the new model is better, replace the old model; otherwise, keep the old model unchanged and adjust the direction of Explore and Exploit[5].

Auto-sklearn integrates classification effect evaluation indicators such as Balanced Accuracy, Average Precision, and Log Loss, but lacks accuracy indicators for binary unbalanced datasets. Due to the particularity of the binary imbalanced data set, the use of existing indicators can easily lead to over-fitting—although the overall accuracy is very high, it cannot make the majority class sample and the minority class sample maintain high accuracy at the same time.

F1-Score can effectively balance the under-fitting and over-fitting problems of minority samples. We modified the logic of the loss function of the binary unbalanced data set: When the binary and unbalanced features of the data set are detected, the loss function can be automatically changed to F1-Score. Every time a new super parameter is introduced into the model, only when the loss value calculated by F1-Score decreases, the super parameter of this round of Bayesian optimization search is set to the current optimal solution. The introduction of F1-Score solves the problem of unclear evaluation of the classification effect of binary imbalanced data sets, so that the model has a better classification effect.

4. Experiments and results

4.1. Performance Comparison before and after framework optimization
In this experiment, variable-controlling method are used to compare the classification accuracy of specified binary unbalanced datasets with the framework before and after optimization. The specific sources and experimental methods of meta-learning and test datasets are as follows.

4.1.1. Experimental method
Before testing the real classification accuracy before and after the optimization framework, it is necessary to import the metadata of the two frameworks. We obtained the meta-learning data features and hyper-parameters of the original auto-sklearn through 24-hour Bayesian optimization. In order to eliminate the interference of training time and machine performance on the performance of the framework, the data set is randomly divided into training set and test set according to 3:2, and Three-fold cross processing is performed. Then, use the framework's respective Bayesian optimization (all using SMAC algorithm) executing for 30 minutes. Finally, the calculated meta-learning features are stored in the local configuration file.

Both the original framework and the new framework use the same test set. Put the dataset into the machine learning framework pipeline, limit the total computing time of each dataset to 10 minutes, and run up to 30 seconds for each Bayesian optimization. In view of the randomness of the model selection each time when the Acquisition Function is used to determine hyper-parameters in the Bayesian optimization process, both frameworks have tested different data sets three times.
4.1.2. Analysis of experimental results

Table 1. Comparative analysis results of the old and new frameworks

| Framework        | Data set ID | Accuracy (Overall) | Accuracy (Minority sample) |
|------------------|-------------|--------------------|----------------------------|
| AUTO-SKLEARN     | 3043        | 97.69%             | 98.27%                     |
|                  | 75092       | 85.27%             | 86.11%                     |
|                  | 75115       | 97.42%             | 69.23%                     |
|                  | 75116       | 98.38%             | 94.23%                     |
|                  | 75117       | 88.39%             | 64.00%                     |
| AUTO-SKLEARN PRO | 3043        | 98.75%             | 97.84%                     |
|                  | 75092       | 91.44%             | 98.35%                     |
|                  | 75115       | 97.35%             | 92.31%                     |
|                  | 75116       | 99.35%             | 98.08%                     |
|                  | 75117       | 92.90%             | 92.00%                     |

According to the experimental result table, after modifying the unbalance dataset matching logic of meta learning of the framework and adding new Bayesian optimization indexes, all the data sets involved in the test can be well classified, and the classification accuracy of the whole, Minority sample and Majority sample has been improved.

4.2. Forest fire prediction based on optimized auto-sklearn

4.2.1. Data source

The data set used in this experiment is UCI Machine learning database, a forest fire data set provided by the University of California at Irvine. The data set records the forest fire occurrence in Montesinho nature park in Portugal for three years from several dimensions such as geographical location, meteorological factors, combustible humidity index, initial spread index and affected area[1]. In the process of data preprocessing, we discretize the labels and transform the regression problem into a binary classification problem.

4.2.2. Experimental method

In each experiment, random unsupervised clustering is carried out according to the geographic location information of the data, and the prediction results of the new framework in two regions are compared with those of general machine learning methods in full samples, to test the regional adaptability of the new framework.

The general classifier selects SVM and random forest that have a better classification effect in the forest fire data set. AUTO-SKLEARN and AUTO-SKLEARN PRO are used to compare the improvement effects of automatic machine learning frameworks, and SVM, RandomForest, AUTO-SKLEARN PRO are used to reflect the regional adaptive capabilities of the framework, so the training data of the automatic machine learning framework uses regional samples instead of full samples.

4.2.3. Analysis of experimental results

Table 2. Comparison results of the models on the forest fire data set

| Forest District | Classifier           | Accuracy (Overall) | Accuracy (Minority sample) |
|-----------------|----------------------|--------------------|----------------------------|
|                 | AUTO-SKLEARN PRO     | 87.3%              | 94.2%                      |
|                 | AUTO-SKLEARN         | 86.4%              | 72.5%                      |
|                 | SVM                  | 84.6%              | 67.3%                      |
|                 | RandomForest         | 75.5%              | 69.2%                      |
| District1       | AUTO-SKLEARN PRO     | 85.6%              | 90.4%                      |
|                 | AUTO-SKLEARN         | 89.5%              | 74.3%                      |
| District2       | AUTO-SKLEARN PRO     | 87.3%              | 94.2%                      |
|                 | AUTO-SKLEARN         | 86.4%              | 72.5%                      |
In the above three forest farm areas, the optimized automatic machine learning framework has significant advantages over the original framework. Compared with the accuracy of general classifiers, the classification accuracy of the optimized automatic machine learning framework is not lower than that of the existing model, and has natural regional adaptability.

5. Conclusion
This paper designs a forestry fire prediction framework based on automatic machine learning, analyzes the principle of the framework, and gives the technical realization points of several core optimization links. This framework can effectively predict the occurrence of forest fires based on meteorological data, and has good regional adaptability and good application prospects. This framework also performs well on other binary imbalanced data sets, and has certain application potential in related fields.

Acknowledgments
This work was supported by National College Students' innovation and entrepreneurship training program (No. G201910022063).

References
[1] Cortez P, Morais A J R. A data mining approach to predict forest fires using meteorological data[J]. 2007.
[2] Feurer M, Klein A, Eggensperger K, et al. Efficient and robust automated machine learning[C]//Advances in neural information processing systems. 2015: 2962-2970.
[3] Feurer M, Klein A, Eggensperger K, et al. Supplementary Material for Efficient and Robust Automated Machine Learning[M]. Advances in Neural Information Processing Systems,2019.
[4] 2018 Kaggle Machine Learning & Data Science Survey[DB].
https://www.kaggle.com/paultimothymooney/2018-kaggle-machine-learning-data-sciencesurvey
[5] Ian D., Michael M., Scott C.Bayesian Optimization Primer[DB].
https://app.sigopt.com/static/pdf/SigOpt_Bayesian_Optimization_Primer.pdf