Thunderstorm weather analysis based on XGBoost algorithm

Lilong Liu[1] *, Huichun Liu [1,],Chaofeng Zhu [2]

[1] College of Surveying and Mapping Geographic Information,Guilin University of Technology,Gulin,Guangxi
[2] College of Software and Internet of Things Engineering,Jiangxi University of Finance and Economics,Nanchang,Jiangxi

KEY WORDS: Thunderstorm weather, ZWD, O_3, XGBoost algorithm, Manual adjustment method, Grid search method

ABSTRACT:

Using the GPS data service platform of China seismological bureau to get the ZTD separated ZWD data pair and the content in the air, and by detecting the O_3 value in the air is an effective method to analyze and study the thunderstorm weather.This paper collected the four foundations of the beibu gulf region GPS station in 10 days in August 2019 data, through ZWD numerical and O_3 values after consolidation, the classification of the training and testing, in XGboost algorithm, manual adjustment method is compared with grid search method, and the results show that the model of manual adjustment method is superior to grid search model and the default model in accuracy and AUC value.

1. GENERAL INSTRUCTIONS

Thunderstorm weather is a kind of meteorological disaster, and cloud lightning is a form of thunderstorm weather, which endangers people's lives and poses a great threat to property safety.At present, many researchers in China use advanced machine learning and data mining to study thunderstorm prediction models.For example,Based on the BPSO-NBayes classifier, the literature studies the lightning prediction technology of the station[1].Cumulonimbus clouds are usually highly developed. In low latitudes, the top of the cumulonimbus cloud can reach 18km, which can reach the height of troposphere. There is a large amount of electric charge on the cumulonimbus cloud.The literature indicates that there is a clear explanation between altitude and lightning current parameters[2].Thunderstorm weather needs a large amount of water vapor as support. Water vapor is positively correlated with precipitatable water vapor, while ZWD is proportional to K of precipitation,The literature has done a good job of this[3] ,It is pointed out in the literature that the tropospheric parameter model is used to forecast the thunderstorm trend[4].However, there are few studies on thunderstorm weather threshold using tropospheric parameters and ozone as reference conditions.When a thunderstorm discharges, some of the oxygen in the air is converted to ozone.By integrating ZTD separated ZWD and detected O_3 values and allocating them to the training set and test set in proportion, the model with high accuracy and good classification effect can be obtained by applying XGboost algorithm and adjusting its own parameters.

2. XGBOOST ALGORITHM

2.1 XGBoost algorithm idea

XGBoost belongs to the category of Boosting algorithm, where the idea is to integrate many weak classifiers together to form a strong classifier.XGboost, on the other hand, is an ascension tree model, so it's also the product of multiple book models integrated into a powerful classifier.The tree model we used is CART regression model.The literature has a good explanation of CART regression model.XGBoost basically grows a tree by constantly splitting features on the observing system, adding one tree at a time, learning a new function and fitting the residual of the previous prediction.All the trees separated from the observation system were trained to predict a sample score (according to the corresponding characteristics of the sample, there will be a corresponding score when each tree falls on the corresponding leaf node), and finally the corresponding score of each tree was summed up.

2.2 XGBoost algorithm principle

$$\hat{y}_i = \sum_j w_j x_{ij} \tag{1}$$

$$f_t(x) = w_q(x), w \in R^T, q : R^d \rightarrow \{1,2,\cdots T\} \tag{2}$$

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{3}$$

Where w_j = the weight of the Jth sample

x_{ij} = the sample corresponding to leaf node

$f_t(x)$ = regression tree

$w_q(x)$ = the score of q on the leaf node

$\hat{y}_i$ = the trees split off by the observation system are summed to give a predicted total score

The core algorithm of XGBoost is:

* Corresponding author: Lilong Liu - email: hn_liulilong@163.com

https://doi.org/10.5194/isprs-archives-XLII-3-W10-261-2020 |

$$Obj = \sum_{i=1}^{n} l(y_i, \overset{\wedge}{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \qquad (4)$$

The first part is used to measure the difference between the predicted score and the real score.In the superposition process of the algorithm, all decision trees need to be taken into account. In order to ensure the improvement of the algorithm, it is necessary to ensure that the current function plus a new function expression, and the overall mean square deviation of the algorithm shows a downward trend.The other part is the regularization term:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \qquad (5)$$

where γ = penalty

T = Node number of leaves

λ = L2 regularization factor for prevent overfitting

As mentioned above, the newly generated tree needs to fit the residual of the previous prediction. When the tree t is generated, the expression of the prediction score is:

$$\overset{\wedge}{y_i}^{(t)} = \overset{\wedge}{y_i}^{(t-1)} + f_t(x_i) \qquad (6)$$

At the same time, we rewrite the objective function as:

$$\iota^{(t)} = \sum_{i=1}^{n} l(y_i, \overset{\wedge}{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t) \qquad (7)$$

Next, we need to find an algorithm that can minimize the objective function. XGBoost's algorithm USES Taylor's second order expansion to approximate it, so the objective function can be approximated as follows:

$$\iota^{(t)} \cong \sum_{i=1}^{n} [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + C \qquad (8)$$

where, g_i is the first derivative,h_i is the second derivative.

For formula (8),It's the same thing as the sum of all the previous models, but it's just a fixed value for formula (8), and we put it in the constant term.And formula (8) can be simplified to formula (9).

$$\tilde{\iota}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)] + \Omega(f_t) + C \qquad (9)$$

It has no effect on the optimization solution. Remove the constant term and sum the loss of each sample of formula (9), and each sample will fall into a leaf and fall into a leaf node. Therefore, we can recombine all the samples in the same leaf node:

$$\begin{aligned} Obj^{(t)} &\cong \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)] + \Omega(f_t) \\ &= \sum_{i=1}^{n} \sum_{i=1}^{n} [g_i w_{q(x_i)} + \frac{1}{2} h_i w^2{}_{q(x_i)}] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w^2{}_j \qquad (10) \\ &= \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda) w^2{}_j] + \gamma T \end{aligned}$$

By rewriting the above equation, the objective function can be rewritten as a unary quadratic function of leaf node fraction, and the optimal sum of objective function values can be solved simply by using the vertex formula directly.we can figure out a optimum value of w and objective function.

$$w_j^* = -\frac{G_j}{H_j + \lambda} \qquad (11)$$

$$Obj = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$

(12)

2.3 Split-node algorithm

Constructing a decision tree based on spatial segmentation is a NP hard problem, and it is impossible to traverse all tree structures. Therefore, XGBoost algorithm uses the same idea as CART regression tree, and uses greedy algorithm to traverse all feature segmentation points of all features, except that the objective function value above is used as the evaluation function.The specific method is that the value of the objective function after splitting is greater than the gain of the objective function of the monad leaf node, and a threshold is added to limit the growth depth of the tree. Only when the gain is greater than the threshold, can the tree split.

2.4 Prediction model based on XGBoost algorithm

In this paper, according to the observation of ZWD and ozone in the air at the time of thunderstorm and non-thunderstorm, the threshold of producing thunderstorms is analyzed and predicted.The flow chart of the design is as follows:

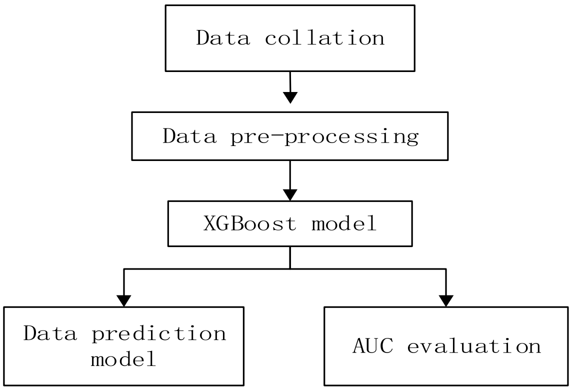

Figure 1.XGBoost prediction model flow chart

2.4.1 Data collation: The data sources in this paper are from the local thunderstorm hours provided by GNSS data products of China earthquake administration,ozone data provided by

https://doi.org/10.5194/isprs-archives-XLII-3-W10-261-2020 | © Authors 2020. CC BY 4.0 License.

Environmental Knowledge Service System, and ZWD data provided by the GPS data service platform of China seismological bureau.The data were collected from August 1 to 10, 2019.In this paper, four ground-based GPS stations are established in Beibu Gulf area of the city,including Nanning,Beihai,Zhanjiang,Haikou.

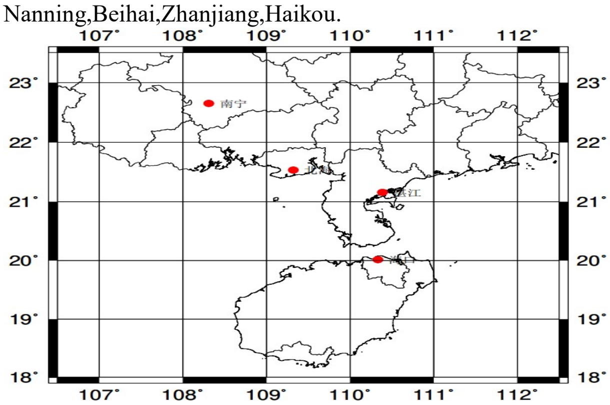

Figure 2.Distribution map of ground-base GPS station

2.4.2 Data pre-processing: All thunderstorm periods of the 4 ground-based GPS stations collected in the Beibu gulf were labeled 1 and 0 for non-thunderstorm periods.In the case of large sample size,

there will be enormous number of randomly selected combinations, and the random combination of different data will also have an impact on the production of results. Therefore, the exhaustion method has no substantive significance in this paper, so it is excluded.In this paper,the ZWD data and ozone detection values are integrated together (each point is in hours) under the premise of considering the time sequence, and distributed to the training set and the test set on a 9:1 ratio.

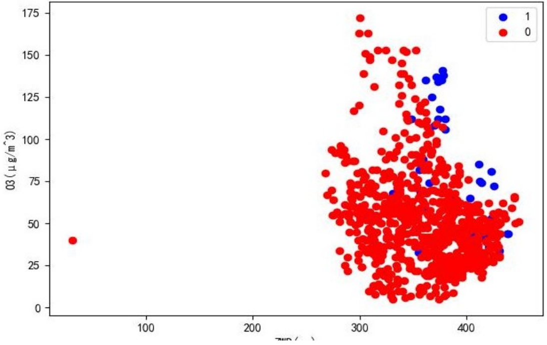

Figure 3. Training set of data on a 9:1 scale

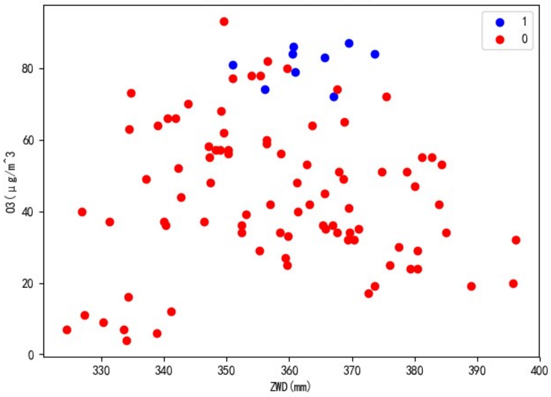

Figure 4. Test set of data on a 9:1 scale

2.4.3 Parameter Optimization: The results of the XGBoost algorithm depend heavily on parameters, including task parameters, general parameters, and auxiliary parameters.Task parameters determine learning tasks and learning objectives, and general parameters determine the type of ascending model.Auxiliary parameters are determined by the ascending model.Starting from the perspective of general parameters, this paper will conduct experiments and analysis on the algorithm of default , manual parameter adjustment method and grid search method in XGBoost algorithm.To optimize the AUC value as the goal of grid search, by setting the parameter range search step size, in the parameter range to find the best parameters.Booster of the two types, gbtree and gbLinear model, the paper default is gbtree.The following table is the general parameters involved in the paper:

Parameter	explanation
min_child_weight	The minimum weight sum of all observations of a subset
gamma	The decrease of the minimum loss function required for node splitting
scale_pos_weigh	Determine the minimum leaf node sample weight sum
max_depth	The maximum depth of each tree
n_estimators	Number of Iterations

Tabel 1.Partial parameter specification table

2.5 Experimental analysis and results

According to the partition of the data in section 2.12, the data is trained and tested. The data is put into the XGBoost algorithm and compared in two ways. The first one is analyzed according to the accuracy rate, and the other one is analyzed according to the AUC value.Firstly,Based on the accuracy, the default model is compared with the model of manual adjustment method as the following figure:

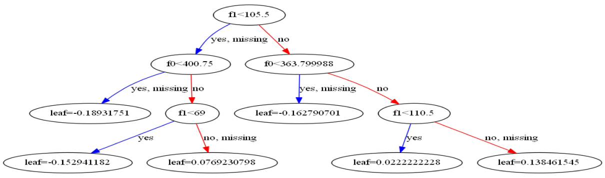

Figure 5.The result of the default model

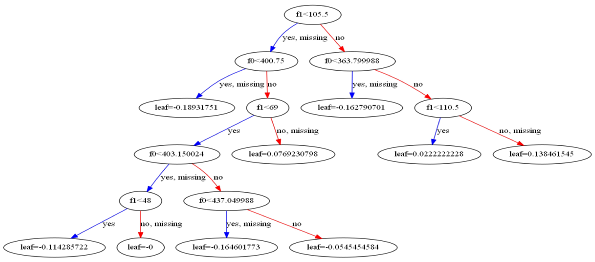

Figure 6.The result of the model of manual adjustment method

Secondly, the optimal parameters obtained by grid search method,according to that parameters,we manually adjusted the parameters around 10 times before taking the average,the comparison between the grid search method and manual adjustment method based on AUC value,as the following figure:

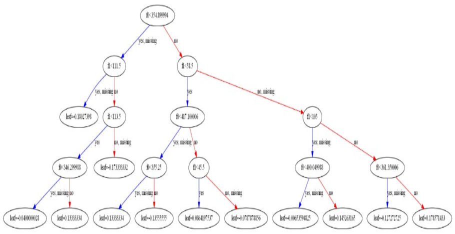

Figure 7.The result of the model of grid search method

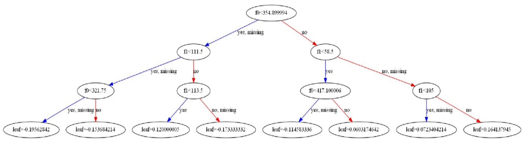

Figure 8.The result of the average after adjusting parameters

	default	grid search method	mannul adjustment method
max_depth	3	4	5
min_child_weight	1	0	1
gamma	0	0.4	0.3
scale_pos_weight	1	8	1
n_estimators	100	52	1000

Tabel 2.Parameter difference table

	Defaults	Manual reference method	Grid Search Method(AUC)	Manual reference method(AUC)
Train Accuracy	93.63%	88.31%	95.02%	97.34%
Test Accuracy	90.62%	88.54%	88.54%	90.62%
Train AUC score	0.79158	0.91031	0.92761	0.77451
Test AUC score	0.54980	0.88697	0.68773	0.5

Tabel 3.Parameter difference table

Above table can be obtained, on the basis of accuracy. The default model training accuracy and precision are all less than the manual tuning method, but it is far less than the latter, the number of iterations in the AUC had lower scores than manual adjustment method, the default model on the prediction precision is superior to manual adjustment method, but the classification effect than manual adjustment method.With the AUC value as the standard, the AUC value of grid search method is far better than the average value of nearby manual adjustment method, and the overall prediction accuracy is 2.2% lower.

ACKNOWLEDGEMENTS

This work was sponsored by the National Natural Foundation of China (41664002; 41704027); Guangxi Natural Science Foundation of China (2018GXNSFAA294045; 2017GXNSFDA198016; 2017GXNSFBA198139); the"Ba Gui Scholars" program of the provincial government of Guangxi;and the Guangxi Key Laboratory of Spatial Information and Geomatics (14-045-24-10;16-380-25-01)

REFERENCES

Bo Zhicheng, Nie Lekui, Sun Dianzhu, Li Yanrui. Node Splitting of Ｒ-tree with Form and Position，Multi-objective[J]. *Modular Machine Tool & Automatic Manufacturing Technique*. 2017.

Cui Yuesheng, Hu Xi. Lightning a ctivity prediction based on IFCM-T-S[J]. *Foreign Electronic Measurement Technology*. 2019.

Cheng Chen，Cheng Xinzhou，Zhang Heng，Han Yuhui. Intelligent Parameter Adjustment XGBOOST and Its Application in Telecom Marketing[J]. *Monthly Focus*. 2018.

Guo Mingang, Gong He. Research on Alex Net Improvement and Optimization Method[J]. *Computer Engineering and Applications*. 2019.

Liu Yajie, Hu Banghui, Wang Xuezhong, Wang Ju, Huang Hong. Research on forecasting technology of thunderstorm interpretation based on BPSO-NBayes[J]. *Journal of the Meteorological Sciences*. 2018.

Li Fen, Xiao Jian, Lin Zhiqiang, Li Zhipeng. Research on BP-ANN Models of Lightning Prediction with Spatio-temporal Characteristics[J]. *Computer and Modernization*. 2019.

Li Yezi, Wang Zhenyou, Zhou Yilu, Han Xiaozhuo. The Improvement and Application of Xgboost Method Based on the Bayesian Optimization[J]. *Journal of Guangdong University of Technology*. 2018.

Wu Ankun, Zhang Yi, Zeng Yong, Wu Shijun, Liu Yun. Effect of Altitude on the Parameters of Lightning Current[J]. *Insulators and Surge Arresters / Insul Surg Arres*. 2016

YE Dezhong, LV Haibing, GAO Yun, BAO Qiuxia, CHEN Mingzi. Novel Real-Time System for Traffic Flow Classification and Prediction[J]. *Special Topic*. 2019.