Analysis of Tourist Flow Forecasting Model Based on Multiple Additive Regression Tree

To cite this article: Chuanli Kang and Junfeng Gu 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **490** 042001

Analysis of Tourist Flow Forecasting Model Based on Multiple Additive Regression Tree

Chuanli Kang[1,2,*] and Junfeng Gu [1,2, a]

[1]Guangxi Key Laboratory of Spatial Information and Geomatics, Guilin 541006, China
[2]College of Geomatics and Geo-information, Guilin University of Technology, Guilin 541006, China

*Corresponding author e-mail: kcl79@163.com,
[a]gujunfeng000@126.com

Abstract. Accurate prediction of tourist flow is a key issue in tourism economic analysis and development planning. This paper proposes a tourist flow prediction model based on Multiple Additive Regression Tree (MART) by using machine learning ideas. The model uses factors such as temperature, sunshine duration, air quality and so on to construct eigenvector, and constructing multiple base learners through the boosting framework to predict tourist flow accurately. Take Guilin city from 2015 to 2018 Tourist as an example for analysis, the prediction accuracy of the model is evaluated by means of the average error, square equalization error and other indicators. The experimental results show that the proposed method has high accuracy in the prediction of tourist flow.

1. Introduction

In recent years, the tourism industry has flourished and played an important role in the national economy. According to the statistics of the national tourism data center, in 2017, the number of domestic tourists reached 50.01 billion, total tourism revenue reached 5.4 trillion RMB, accounting for 11.04% of the total GDP. In order to make a response strategy to the tourism situation, the scientific and accurate prediction of tourist flow has become a hot research topic nowadays.

The existing forecasting methods are mainly divided into two types [1]. One is a statistical forecasting method, such as exponential smoothing, ARIMA, Kalman filtering, etc. The other is a machine learning forecasting method, such as BP neural network, decision tree, support vector machine (SVM), etc. Scholars at home and abroad have studied these two methods in the prediction of tourist flow.

This paper put forward a kind of tourist flow prediction model based on Multiple Additive Regression Tree (MART) by using ensemble learning boosting framework, the model consists of multiple regression trees, each subtree is learning at a certain rate in the negative gradient direction of the residual error of the previous tree, finally, the model is predicted by linear combination of multiple subtrees. It is proved by experiments that this model has high precision and obvious superiority in the prediction results of tourist flow.

2. The algorithm principle of MART

2.1. Regression Tree

The basic learner of the MART model is the CART regression tree [2], the tree generation method is to divide the eigenvector space into branches, select each threshold value of each eigenvector value, find the branch basis by minimizing the mean variance, and finally meet the preset termination condition. Assuming a regression tree has n characters, each character has different values, through the exhaustive each characteristics of each selection on space division, until get the characteristic value, make the loss function is minimal:

$$\min_{js}[\min_{c_1} Loss(y_i, c_i) + \min_{c_2} Loss(y_2, c_2)] \tag{1}$$

If the input space is divided into n units: $R_1, R_2, \dots, R_n$,then the output value of each region is the average value of y value of all points in the region:

$$C_n = ave(y_i \mid x_i \in R_n) \tag{2}$$

The resulting regression tree as:

$$f(x) = \sum_{n=1}^{N} C_n I(x \in R_n) \tag{3}$$

2.2. Multiple Additive Regression Tree

MART generates strong learners in the form of a weak learner set, and then makes learning and model prediction of training samples. The core thought is to add a new regression tree to minimize loss function in each iteration, each new tree is conducted on the residual error of learning in a tree, and all of the negative gradient direction along the loss function training, through multiple iterations, training more weak learning, eventually these weak learning linear combination to create a strong learning. It is assumed that the sub-tree of the prediction model basic learner is:

$$f(x; \{R_j, b_j\}_1^J) = \sum_{i=1}^{J} b_j \tag{4}$$

In an expression, $\{R_j\}_1^J$ is subtree segment space, $\{\mathrm{b}_j\}_1^J$ is the function value of subtree f_x on segment space $\{R_j\}_1^J$, and J is the number of leaf nodes. The algorithm principle of MART flow as follow [3]:

① Initializes f_0 and the robustness of MART makes it insensitive to the selection of initial values.

② Solve for response $\tilde{y}_i$,which is a variable positively related to residual: $y_i - F_{k-1}(x_i)$ and when the loss function is $Loss$=L$(y_i, f(x_i))$, the derivative of the loss function L with respect to $F(x)$ is gradient g_i

$$g_i = \frac{\partial L(y_i, F(x_i))}{\partial F} \tag{5}$$

③ The residual is calculated according to the loss function, and the region is segmented by learning the residual and fitting the training sample $\{(x_i, y_i)\}_1^N$.

$$\{R_j, b_j\}_1^{J^*} = arcmin_{\{R_j, b_j\}_1^J} \sum_{i=1}^{N} [\tilde{y}_i - f_k(x_i; \{R_j, b_j\}_1^J)]^2 \tag{6}$$

④ In order to avoid over-fitting, each regression tree is multiplied by a certain learning rate to scale the step length of the gradient descent process, after residual learning finished basic learning subtree f_k as:

$$f_k = \rho * f_k = [arcmin_\rho \sum_{i=1}^{N} L(y_i, F_{k-1}(x_i) + \rho f_k(x_i))] f_k \tag{7}$$

⑤ Each training base produced by learning subtree f_k stack to the base learning integration tree F_k, repeat steps ①-④until the loss function square error tends to set threshold or convergence, F_k at this time for each iteration of the set of f_k model, the base learning integration tree as follows:

$$F_k = F_{K-1} + f_k = \sum_{i=0}^{k} f_i(x) \tag{8}$$

As an ensemble learning method, the learning effect of MART is better than that of other weak learners, and its generalization ability is stronger.

3. Model Specification

3.1. Data pre-processing

Guilin city is a famous tourist city in China and it is of practical significance to take Guilin city as a case study for tourists flow prediction. In order to verify the accuracy of the model and the rationality of the algorithm, this paper took the tourist flow from January 2015 to June 2018 as the research sample. Due to the large monthly tourist flow value, when the gradient drops, the direction of the gradient will deviate from the direction of the minimum value, which is easy to cause the loss of precision. To Min-Max normalization the data for this purpose. The raw data is preprocessed according to equation (9), and the result is mapped between [0-1].

$$\frac{x_i - x_{\min}}{x_{\max} - x_{\min}} 0.7 + 0.15 \tag{9}$$

3.2. Construct the eigenvector

There are many factors influencing tourist flow, so the selection of eigenvector must be characterized. Pearson correlation coefficient was used to evaluate the correlation between influencing factors and tourist flow. According to equation (10), the correlation coefficient between each variable and tourist flow is obtained by using network crawler technology. The results are shown in table 1.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{10}$$

Table 1 Correlation coefficient

influencing	correlation	influencing	correlation	influencing	correlation
high temperature	0.5900	AQI index	-0.4927	PM2.5	-0.4078
low temperature	0.5949	precipitation	-0.2154	sunshine duration	0.5187
relative humidity	-0.2598	rainfall days	-0.2303	wind speed	-0.1123

According to the principle of Pearson correlation coefficient, the correlation coefficient of the correlation of absolute value is bigger and stronger, as shown in table 1, select month average temperature, monthly mean temperature, monthly mean sunshine duration, monthly average AQI index and monthly average concentrations of PM2.5 as tourist flow forecast model relevant variables and characterizing the above variables to construct the eigenvector.

3.3. Prediction model construction

Since the MART prediction model cannot directly deal with the eigenvector with high complexity, it uses part of the eigenvector training MART model firstly, then constructs new eigenvector with the trees learned by the MART model, and finally adds these new eigenvectors to the original eigenvector training model [4]. According to the base learning subtree segmentation of residual after learning for the first time, fitting the training sample. For region segmentation, eigenvector using the grid search method, make each iteration of f_k vector to specify the study, gradually approaching loss function minimum, then the solution of vector in response to the residual. On this basis, f_k generated by each training is added to F_k. Repeat the above steps until the loss function squared error reaches the set threshold or tends to converge. At this point, F_k is the model set of f_k generated by each iteration, besides, the average sunshine duration eigenvector training is completed and the average sunshine duration eigenvector subtree is constructed. For the learning of other eigenvectors, as mentioned above, each eigen learning, a new eigenvector is constructed on the eigen subtree of the eigen vector basis, After the completion of all eigenvector learning, the integrated tree of prediction model is obtained, the training test is carried out on the total training sample, and the complete travel passenger flow prediction model is constructed

4. Experiment simulation

In this paper, the tourist flow of Guilin from January 2015 to June 2018 was taken as the research sample, and the cross-validation method was adopted to make all the data used for learning and testing and predict the standardized tourist flow. Comparison between the prediction model and the real value of MART tourist flow is shown in figure 1. Average error E_{ME} , mean square error E_{RMSE} and average absolute error percentage E_{MAPE} were used to evaluate the prediction effect of the model, as:

$$E_{ME} = \frac{1}{\mathrm{n}}\sum_{1}^{\mathrm{n}}[f(i) - y(i)] \tag{11}$$

$$E_{RMSE} = \sqrt{\frac{\sum_{1}^{n}[f(i) - y(i)]^2}{n}} \tag{12}$$

$$E_{MAPE} = \frac{1}{n}\sum_{1}^{n}\left|\frac{f(i) - y(i)}{y(i)}\right| *100 \tag{13}$$

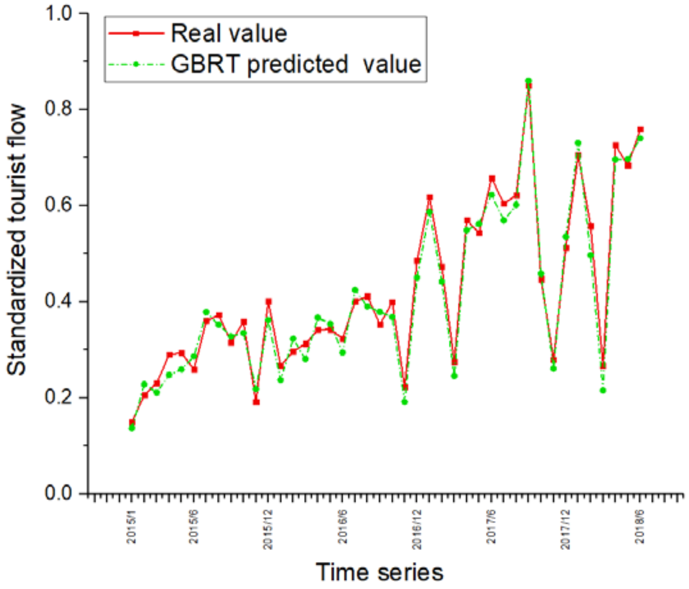

Fig. 1 Comparison between prediction results based on MART and the real value

According to calculations of equations (11), (12) and (13), the average error E_{ME} = -0.01077, mean square error E_{RMSE} =0.1283 and average absolute error percentage E_{MAPE} =7.38% based on the MART travel flow prediction model can be obtained. From the perspective of square mean square error E_{RMSE} ,the overall deviation between the tourist flow predicted by this method and the actual tourist flow is lower, and the prediction results are better. It can be seen from Fig.1 that the tourist flow prediction model based on MART can accurately predict the tourist flow and is an effective tourist flow prediction and evaluation model.

5. Conclusion

This paper proposes a tourism flow prediction model based on Multiple Additive Regression Tree (MART) by using ensemble learning boosting framework. The prediction accuracy of this model is still high under relatively small parameter adjustment, and it is not sensitive to the selection of initial value. Taking Guilin city as an example, the prediction model of this paper is verified experimentally, the experimental results show that the tourist flow prediction model in this paper is accurate and has certain practical application value in the development and planning of tourism economy.

Acknowledgement

In this paper, the research was sponsored by National Natural Science Foundation of China (Project No.41461089) and BaGui Scholars Post of Special Project of Guangxi(Project No.163802515)and Key Laboratory of Spatial Information and Geomatics of Guangxi(Project No. 151400720)

References

[1] L Zheng,F Zhu,A Mohanmmed.Attribute and Global Boosting: A Rating Prediction Method in Context-Aware Recommendation[J].Computer Journal,2018,60(7):957-968.

[2] Friedman J H. Greedy function approximation: a gradient boosting machine[J], Annals of Statistics,2001,1189-1232.

[3] Ran-Ran He, Yuanfang Chen, Qin Huang. Evaluation of ocean-atmospheric indices as predictors for summer streamflow of the Yangtze River based on ROC analysis[J].Stochastic Environmental Research and Risk Assessment. 2018,32(7):1903–1918.

[4] Chao Fan,Diwei Liu,Rui Huang,el al. PredRSA: a gradient boosted regression trees approach for predicting protein solvent accessibility[J].BMC Bioinformatics.2016,17 Suppl 1(S1):85-95