2017 10th International Symposium on Computational Intelligence and Design

Prediction of Road Traffic Congestion Based on Random Forest

Yunxiang Liu[1], Hao Wu[1]
1.School of Computer Science and Information Engineering
Shanghai Institute of Technology
Shanghai, China
E-mail: 947538920@qq.com, yxliu@sit.edu.cn

***Abstract*—With the process of urban modernization becoming faster and faster, there are more and more vehicles in the city, and the situation of urban traffic congestion is becoming more and more serious. In this paper, a model of traffic congestion prediction is constructed by using machine learning classification algorithm - random forest to construct traffic congestion state prediction model. The random forest algorithm has the characteristics of high robustness, high performance and high practicability. The weather conditions, time period, special conditions of road, road quality and holiday are used as model input variables to establish road traffic forecasting model. Finally, the results show that the traffic prediction model established by using the random forest classification algorithm has a prediction accuracy of 87.5%, and the generalization error is low, and it can be effectively predicted. Moreover, the calculation speed is fast, and it has stronger applicability to the prediction of congested condition.**

Keywords-random forest; decision tree algorithm; traffic congestion; prediction accuracy

I. Introduction

With the rapid growth of the urban population, more and more urban vehicles are being carried out. The process of modern urbanization is speeding up. The urban road is becoming more and more complicated, and the urban traffic problems are becoming more and more serious. And road traffic congestion [1]-[2] is one of them. In large cities, when traffic congestion occurs, if it is not handled in time, it will lead to more and more crowded areas and can even lead to traffic paralysis. The first thing to tackle traffic congestion is to prevent it from happening. Therefore, the establishment of a traffic congestion effective forecast is conducive to the preparation of targeted preventive measures and early warning. In order to predict the problem of traffic congestion, domestic and foreign scholars have made some good progress in carrying out forecasting research. Scholars from different disciplines and fields have analyzed from different perspectives, time series method, nonparametric regression method[3] and neural network method are applied to road traffic congestion prediction. These methods provide a good effect for traffic congestion prediction, and can establish a rapid and effective forecasting model. Most of these methods are the prediction and analysis of traffic flow parameters. Based on the traffic state data corresponding to different traffic conditions accumulated in the previous traffic platform, this paper establishes the traffic congestion prediction model based on random forest[4]-[6] to predict traffic congestion.

II. Research Methods

A. *The principle of random forest algorithm*

The random forest algorithm was first proposed in 2001 by Leo Breiman, a professor at the University of California, Berkeley, in the United States. The algorithm is a supervised data mining algorithm. The algorithm is a supervised data mining algorithm. Random forest is a classifier that utilizes a large number of CART decision trees [7]-[9]. So this paper first introduces the CART decision tree.

The CART algorithm is a single classifier, which calculates the GINI index of all the properties of the current sample set, and sorts all the properties of the GINI index,the minimum attribute of the GINI index is the root node of the CART decision tree, and then the sample set is divided into two parts by the GINI index of the attribute.In the process of building the CART decision tree, make full use of the binary tree, in the sub-subset after the recursive repeat this action, so that the final generation of non-leaf nodes have left and right branches, until all the leaves in the node samples Of the categories are roughly the same, or the next subsequent split attribute has been gone.

The GINI index is the impure degree of the metric data partition E. Defined as follows:

$$\text{GINI（E）} = 1 - \sum_{i=1}^{m} p_i^2 \tag{1}$$

Among them, class sets $\{C_1, C_2, \ldots C_n\}$, $| C_i |$ is the number of samples belonging to C_i in E, and $p_i = | C_i | / | E |$ is the probability of class C_i in E.

When the attribute A divides the training sample set E into E_1 and E_2, then the GINI index formula of E after separation is:

$$\text{GINI}_A(E) = \frac{|E_1|}{|E|} GINI(E_1) + \frac{|E_2|}{|E|} GINI(E_2) \tag{2}$$

Among them, $| E_k | / | E |$ is the probability of a subset of k (k = 1, 2).

Random forest is a combinatorial classifier model formed by many unpruned CART classification trees{h (x, Θ_k) k = 1, 2,... k}(x is the input variable; Θ_k is the formation of independent indentically distributed random vector).The first randomized idea is to create k random vectors Θ_1, Θ_2 ... Θ_k according to the bootstrap re-sampling, and then turn each random vector Θ_i into a non-pruning decision tree to

2473-3547/17 $31.00 © 2017 IEEE
DOI 10.1109/ISCID.2017.216

obtain the k tree decision tree{ $h_1(x)$,$h_2(x)$ … $h_k(x)$}, there is no association between each generated decision tree. Second randomization thought is in the generation of decision tree, select the attribute is randomly generated, in all probability, such as the properties of the concentrated, select feature attribute values constitute feature attribute subset, reuse of these attributes subset feature attribute to the decision tree. The large number of decision trees formed together are called random forest(RF).

Assuming *y* is the output variable, the sample data set consisting of (*x, y*) is called the original sample data set. Classification result by the sequence of the last we need all the decision tree classification results in the comprehensive decision, and this was voted the most simple method is adopted, the category of the input variables *x* is the category of the most votes. The final classification results are as follows:

$$\mathrm{H}(x) = \arg \quad \max \sum_{i}^{k} I(h_i(x) = \mathrm{y}) \qquad (3)$$

Among them: H (*x*) combined classifier model, hi as a model of a decision tree classification, I(•) as the indicator function (shown in sexual function refers to the function make the collection has the value is 1, not is 0), y is the output variable.

B. *Establishment of random forest algorithm model*

The process of establishing a random forest algorithm model is as follows:

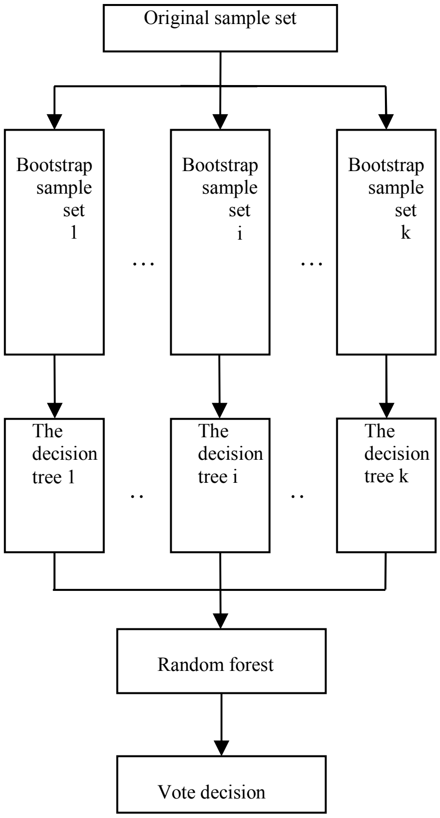

Figure 1. Generation process of random Forest

Step 1 From the original data using the bootstrap method to select the k different sample set of data, each bootstrap extracted by the sample set is the decision tree of all the training data, and the number of samples per sample set is equal to the original data set.

Step 2 The *k* sample sets selected by the bootstrap method are used to construct *k* unrouted decision trees. In the process of generating each decision tree, in order to generate the nodes of the decision tree, we need to select *m* attribute attributes ($m \leqslant \mathrm{M}$) from the M attribute of the original data set as the candidate feature attribute. The decision tree is constructed by using the m candidate attribute attributes randomly selected, and the complete growth of the pruning trees is not carried out. Each tree of decision tree is classified as a complete decision tree, finally get k classification results.

Step 3 Finally, the way to vote, according to the results obtained k classification, the final classification of the output variable to vote to get the most votes of the type of output variables for the final category. The model is shown in Figure 1.

C. *OOB estimates and the importance of attribute variables*

As a result of the bootsrap re-sampling method to generate k data sets, nearly 37% of the samples in the original data may not be selected, and these unselected samples are called Out-Of-Bag (OOB) data. Every tree of the random forest model will have a OOB error estimation, all the trees have OOB error estimates averaged as a generalization of the model error estimation, it is used to test classification performance of the model. A large number of experiments show that the error of OOB is not very different from that of cross-validation if the number of trees is large enough. For the generated stochastic forest model, the noise is added to one of the characteristic attributes, and the different OOB accuracy before and after the noise is obtained is used to test the performance of the model. The larger the reduction of the value is, the more useful the characteristic attribute is.

III. Example's Application

Traffic congestion is a common problem in road traffic. When the number of vehicles exceeds the upper limit of the road, traffic congestion can easily occur. Environmental factors affecting traffic congestion are: weather conditions, different time periods, special road conditions, road quality, holidays.

A. *Measurement standards affecting traffic congestion status*

In this paper, traffic *Q* and road length *L* are used to measure congestion by influencing traffic congestion. Through these two measures, the driving time, average speed, travel speed, delay, total delay and congestion index *CI* can be obtained directly or indirectly. Traffic congestion can be divided into smooth, congestion, blockage, can be used to measure the congestion index. The congestion index (*CI*) is calculated as follows:

$$CI=\frac{t_L-t_0}{t_0}$$

(4)

In the formula: tL is the driving time of road sections; t_0 is the travel time in free flowing condition.

CI is dimensionless index, according to the research data. When $0 \leqslant CI \leqslant 0.15$, the traffic state is defined as smooth; $0.15 < CI \leqslant 0.8$, the traffic state is defined as congestion; $CI > 0.8$, the traffic state is defined as blockage.

B. *Attribute selection of traffic congestion prediction model*

The final output category of the model is the congestion state, which is divided into smooth, congestion and blockage. The properties of traffic data are shown in Table 1.

In this paper, we obtain 1124 data from different sections of Shanghai traffic management information department, and construct the original data set with the selected attributes and the crowded state of *CI*.

TABLE I. ELECTION OF DATA’S ATTRIBUTE

Name of attribute	the value of the attribute
Weather	sunny、foggy、rainy、snowy
Time	peak_hour、non_peak_hour
Holiday	yes、no
Special_condition	false、true
Quality of road	ok、bad

C. *Evaluation of model classification performance*

In this paper, the classification performance of RF model is evaluated by using the overall classification accuracy (A_{CC}) index. The A_{CC} index indicates the ratio between the predicted value of the model and its real value, and the larger the ratio, the better the classification performance of the model. The calculation formula is as follows:

$$A_{CC} = T_P / T_N$$

(5)

In the formula, T_P is the correct classification of samples; T_N is the total sample size.

D. *Construction of random forest model*

In this paper, the randomForest package in R is selected to construct the traffic congestion model based on RF algorithm. Setting up this prediction model requires setting two parameters, ntree and mtry. Among them, ntree represents the number of trees [10]. The larger the value of the tree, the smaller the fitting effect is, and the value of ntree is often set to 100, and the relationship between OOB error and ntree size can be obtained by calculation, as shown in figure 2. While mtry indicates the number of feature attributes to be selected, its value is generally the square root of all characteristic attributes, and the number of feature attributes of this article is 5, so the value of mtry is 2.

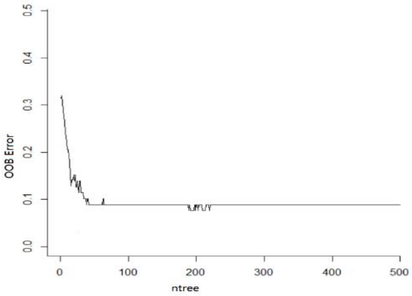

Figure 2. Relationship diagram of ntree and OOB error

Figure 2 shows that when ntree = 100 or so when the OOB error tends to be stable, its value is also smaller, then the random forest model classification performance is higher. So we set the value of ntree to 100, mtry value is set to 2, set these two parameters, you can train the original data set of the first 1100 data and get the desired random forest congestion state prediction model. The remaining 24 sets of data as test data into the random forest model, the 24 sets of data classification to determine the results shown in Figure 3. In the graph, 0 in the vertical coordinate represents smooth, 1 represents the congestion, and 2 represents blockage. As shown in Figure 3, the final classification accuracy rate of 21/24, the classification of the correct rate of comparison is high. Indicating that the random forest traffic congestion prediction model is effective and can be applied.

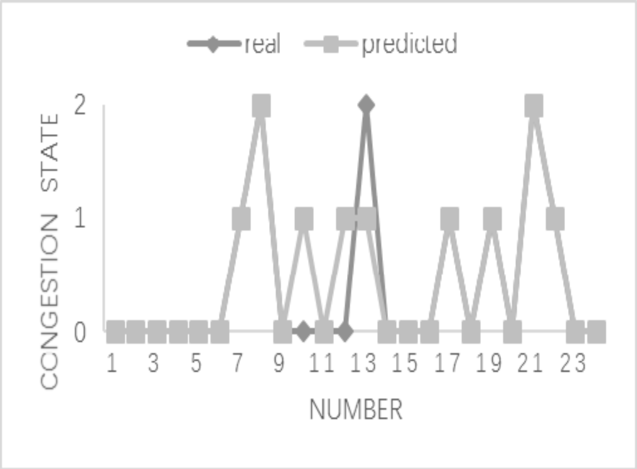

Figure 3. Comparison of traffic congestion prediction

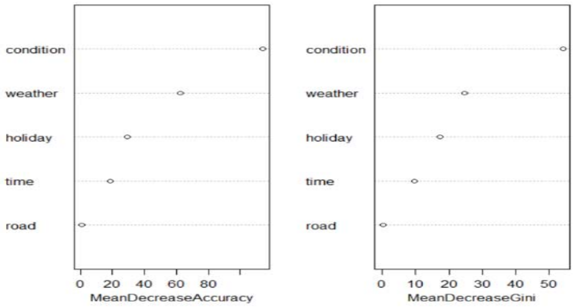

Figure 4. The importance of influencing factors

In addition, the use of the stochastic forest prediction model can also compare the relative importance of environmental factors affecting the congestion, and the importance of the environmental factors affecting the congested state can be obtained. The results are shown in Figure 4.

IV. CONCLUSION

Scientific correct prediction of traffic congestion can ensure a safe traffic environment, and can be very good to prevent traffic congestion, to avoid traffic accidents. The random forest model does not need to set the weight of the property first, how to classify it, the calculation process is simple and the calculation quantity is small. The model needs to be set with fewer parameters, suitable for wide range of platforms, simple and easy to implement, and is a fast and effective data mining model. The final test results show that the prediction model established by RF algorithm can effectively predict the traffic congestion and can also find the relative important environmental factors that affect the traffic congestion. This paper presents an effective method for traffic congestion forecasting and provides effective scientific basis for traffic management to prevent traffic congestion.

ACKNOWLEDGMENT

The research is supported by National Natural Science Foundation of China (No.41671402;No. 61401281).

REFERENCES

[1] Gu Jiuchun, Yao Chen, Liu Lu, “Urban road traffic congestion identification based on multi-attribute decision-making,” Control Engineering, vol. 23, no. 4, pp. 502-505, 2013.

[2] Liao Ruihui, Zhou Jing, “A traffic congestion warning model based on cloud support vector machine,” System Engineering, vol. 33, no. 4, pp. 149-153, 2015.

[3] Li Cheng, Yang Shuyuan, “Neural network 70 years: review and prospect,” Journal of computer science, vol. 39, no. 8, pp. 1697-1716, 2016.

[4] Li Xinhai, “Application of random forest model in classification and regression analysis,” Journal of Entomology, vol.50, no.4, pp.1190-1197, 2013.

[5] Dong Shishi, “Analysis of random forest theory,” Integration Technology, vol. 1, no. 2, pp. 1–7, 2013.

[6] Hu Tianyi, Dai Bo, “The slope stability prediction model based on random forest classification algorithm,” People's Yellow River, vol. 39, no.5,pp. 115–118, 2016

[7] Pan Dasheng, An improved ID3 decision tree mining algorithm,” Journal of overseas Chinese university (natural science edition), vol.31, no.7, pp.71-73, 2016.

[8] Xie Niuniu. “Decision tree algorithm overview,” Software guide, vol. 14, no. 11, pp. 63-65, 2015.

[9] Zhang Liang, Ning Qian, “Two improvements and applications of CART decision tree,” Computer engineering and application, vol. 36, no. 5, pp. 1209-1213, 2015.

[10] Liu Min, Lang Rong, “he number of trees in random forest,” Computer engineering and application, vol. 51, no. 5, pp. 126-131, 2015.