AN APPLICATION OF RANDOM FOREST METHOD ON TRAFFIC BIG DATA

Waseda University Regular Member ○Kuniaki SASAKI
Waseda University Student Member ○LIU Xingwei

1. INTRODUCTION

Random forest is regarded as the powerful supervised machine learning algorithms, and it is both the excellent prediction and classifier method. Random forest method is also called random decision forest which can create forests with many decision trees. It is first proposed by Ho (1995) to do classification through handwritten digits. Consider random forest has the advantages of highly accurate and versatility, Beriman (2001) applicated the random forest to optimize model and do prediction. Random forest method is a useful tool on both prediction and classification, it has been applied well on many fields, such as product recommendation, customer segmentation, risk detection and so on. In the recently, with the development of ITS, we already obtained traffic big data, including smart phone data, personal travel data, traffic flow data which obtain lot of information which could be dig deeply. The following content aims to explain the principle of random forest method, the analysis processing of random forest and explore how it applies on big traffic data.

2. THE PRINCIPLE OF RANDOM FOREST METHOD

The basis of the random forest method is the decision tree which we would provide the overview of the decision tree first. The decision tree is a typical inductive approach, and it is a flowchart with the tree structure as name suggest. Each node in a tree represents the different feature, and each branch represents their performance for their features. Through tree structure, one dataset (training dataset) can be separated into subsets based on vary variables itself. Even though decision tree method has the strengths of generating the understandable rules and fast calculation speed, decision tree may have errors when dataset is less or situation is complicated. In order to broaden the range of application, large number of individual decision trees can be combined to decide the final results (shown in Fig.1) which is basis principle for the random forest method. The random forest method employs the bootstrapping to generate the training data and trees will be produced according to the different sub-data. Besides bootstrapping, the features are also selected randomly. That means in each random forest, trees are trained by different data and features. It is unlikely to obtain poor accuracy which can lead to better stability. Due to the large number of trainings with bootstrapping, the algorithm is hard to overfitting as well.

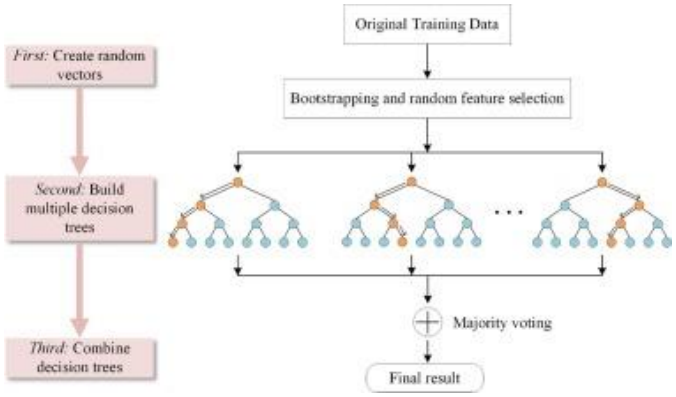

Fig.1 Illustration of the random forest algorithm. Source: Cheng et al. (2019)

Keywords: Random forest method, Decision tree, Big traffic data, Application
Contact address: Ohkubo 3-4-1, Shinjuku-ku, Tokyo, 169-8555, Japan, Tel:+81-3-5286-3398
E-mail: xingwei_18@fuji.waseda.jp

3. THE ANALYSIS PROCESSING OF RANDOM FOREST AND APPLICATION

The setting of hyperparameters plays an important role on the performance of the random forest method and its calculation efficiency. Three main parameters need to be tuned carefully:

(1) The number of trees, n. It specifies how many decision trees will be built in this random forest method.

(2) The maximum depth of trees, d. It represents the maximum of levels in each decision tree.

(3) The number of splitting variables, m. This parameter restricts the partitions for branches in one tree.

Proper parameters can expand the forest size, improve the performance of each individual tree and reduce the crossover among the forest.

What's more, one of the key results for random forest method is the relative importance which explains random forest can explore the relationship between the explanatory variables and predicted value without assuming priorly. The properties of random forest can help us to figure out the complicated relationship among big traffic data. As for the built environment variable, it is summarized as 5Ds, including density, diversity, design, distance to transit and destination accessibility. These five factors are reflected in the data could be as following: population density, streetscape greenery, road density, traffic flow value, congestion situation, travel time, travel purpose, land use mix (entropy), number of bus/subway station, the distance to the nearest square / park / café / restaurant and so on which is convenient to obtain with the help of traffic big data. Besides, these variables of built environment, as well as other socio-demographics variables including age, gender, income, education level, house size, family members, can be regarded as the explanatory variables. These explanatory variables are the input for random forest method. Following tuning and verification mentioned above, the relative importance of variables and partial dependence plots will be produced which should be analyzed with caution. We could find variables caused the different influences on travel mode choice. Subsequently, we can further analyze the travel mode choices based on age, gender as well as other individual attributes.

We take the PT data as an example. The latest PT data collected the person trip from September to November in 2018 on weekdays in Greater Tokyo Area that includes the travel diary for 24 hours for participants. we focus on the travel mode choice from elder people in Chiba city, Chiba prefecture considering 26.1% of resident were aged 65 or over. Besides, among elder people, the car ownership for male is around 76%, for female is 62%, the bicycle ownership for male is around 65% and 58% for female. Almost 79% male still have driving license, but this rate is only 43% for female. We can image that elder male and female have different travel habits which may cause the different travel mode choice. The next step is to explore the relationship between their travel mode choice and other variables counted from PT data, including occupation status, household income, population density and so on. In this part, random forest can rank the relative importance of explanatory variables and will perform well. After knowing the relationship between elder people travel mode choice and socio-demographics, we can make some measures and provide more convenient transportation to help elder people, especially for whom living alone. This result will offer a guidance to build an elder-friendly society.

REFERENCE

Long C, Xuewu C, Jonas D V, Xinjun L, Frank W.: Applying a random forest method approach to model travel mode choice behavior, Travel Behaviour and Society, 2019, 14: 1-10

Yang, L,Yibin A, Jintao K, Yi L, Yuan L.: To walk or not to walk? Examining non-linear effects of streetscape greenery on walking propensity of older adults. Journal of transport geography. 2021, vol. 94.

Breiman L.: Random forests. Machine Learning, 2001, 45: 5-32

Ho T K.: Random decision forests. In: Proceedings of 3rd international conference o document analysis and recognition. IEEE, 1995, pp.278-282