**Prediction Model for Road Traffic Accident Based on Random Forest**

Rong CHENG, Meng-Meng ZHANG and Xue-Mei YU  
Fundamental Institute of Beijing Polytechnic, Beijing, China

**Keywords:** Traffic Accident, Prediction Model, Random Forest

**Abstract.** More than 1.3 million people worldwide die in road traffic accidents every year, and the number of traffic accident deaths in China is increasing by 10% every year. Traffic safety has become an important direction for the development of road transportation systems. At present, research in progress of road traffic accidents is mainly from accidents numbers and the degree of accident damage. Firstly, this paper proposes a random forest prediction model for the prediction of the number of road traffic accidents, which applied to the actual to verify the validity of the model by using the R software to solve it.

1. **Predictive Model**

   It researched on the amount of the traffic accidents, which the traditional statistical models commonly included that linear regression models, poisson regression models, negative binomial distribution models, zero-inflated model and hurdle models. The following elaborated Random Forest expressions, features and in which using condition.

2. **Random Forest Prediction Model**

   The random forest method was first proposed by Leo Breiman and Adele Cutler. It is a classification algorithm that trains and analyzes samples through multiple trees. Intuitively, the decision trees which are classification models that predict categories or taking label by by inputting features. While the random forest is a model composed which multiple decision trees due to different training data. Its result is the combination of all decision tree outputs, which is more stable and with higher prediction accuracy. Taking randomly sampling with replacement in the Random forests progress, and select each tree from the total set. The relationship of the feature at this node needs to be calculated at each node, and then select one branch to move to the next node according to the operation result. We extract the use features randomly that sampling without replacement when training the nodes of each tree by Leo Breiman's proposal. This ratio may be assumed to be when S is the total number of features, then taking sqrt(S), 1/2sqrt(S), 2sqrt(S)

   The training process of random forests can be summarized as follows:

   First, for the symbol introduction: Let's assume that the training set is E, the test set T, the feature dimension is A, a is the number of CARTs used, d is the depth of each tree, b is the feature quantity used by the node; n is the node The minimum number of samples, l is the minimum conditional increase on the node.

   For any tree, it may be set as No. i(i=1-a): the training set E (i) as large as E is extracted from E which is as the sample of the root node ,and training started from the root node; if the termination condition is reached at the current training node, which is set as leaf node. If it is not reached, then the b-dimensional feature is randomly selected from the A-dimensional features without replacement, then observe that the one-dimensional feature g and its threshold th with best classification results, if the sample with the g-th dimension of the sample greater than or equal to th then divided into the right node, then the rest is divided into the left node. The other nodes are keeping training. Repeat the above until all nodes have been trained or marked as leaf nodes; finally repeat the previous steps until all CART has been trained.

   The prediction process using random forests is roughly as follows:

   For any tree, it may be set as the i-th (i=1-a), then: the root node of the current tree as the starting position, and enter into whether the left node (<th) or the right node (>= Th) that judging by
threshold \( t_h \), then until leaf node, output the predicted value. Then repeat the above steps until all trees output the predicted value.

3. Model Application

This paper analyzes the road traffic accident data what a certain section Florida in the United States by R language programming. The number of traffic accidents is taken as the dependent variable, the length of the road segment, and the traffic volume, the road surface condition are used as independent variables (model parameters are defined as Table 1 shows the establishment of a random forest regression model. By ranking the importance of the independent variables, that selected the two most important factors affecting which are the length of the road segment and the traffic volume. Re-establishing a regression model.

| Name                | Definition       |
|---------------------|------------------|
| Dependent variable  | accident         |
| Count               |                  |
| Continuous dependent variable | road segment |
| AADT                | Year             |
| PSR                 | pavement         |
| AVGTR               | traffic          |

| Categorical Independent Variable | Definition   |
|----------------------------------|--------------|
| NOLA single side                 |              |
| NOLA single side                 |              |
| RURAL Country Road               |              |
| RURAL Urban Road                 |              |

4. Random Forest Model

It is easy to operate that a ready-made random forest bagging (R) in R. The data is divided into two parts in this instance: 70% of the training samples and 30% of the testing samples. The first time all the independent variables are established (Including the continuous variables and categorical variables) as a function of the number of traffic accidents (continuous variables), which the interpretable variance is 17.87%, that the specific gravity. It is acceptable that the proportion analyzed the number of traffic accidents. Making important ordering analysis of each dependent variable, the results showed that the two indicators are consistent (as shown in Figure 1), so the AADT and SLENGTH factors are highly interpretable for the number of traffic accidents. We use principal component analysis to the importance of the independent variables was screened (as shown in Figure 2), and the results were consistent with those of the random forest method.

Figure 1. Random Forest Importance Ranking Results.
The most important two dependent variables (AADT, SLENGTH) were chosen to establish a regression model with the dependent variable (Count). The results of the random forest showed an interpretable variance of 22.34%, which was about 5% higher than before. It is verified that the established model (including all feature vectors and key feature vectors) by using test data. The verification results show that the model built by the key features has better fitting effect and is higher than the regression model established by the full feature. The specific results are shown in Figure 3.

5. Summary and Prospect

The random forest has a good predictive effect on the number of road traffic accidents that according to the above actual case analysis, certainly, the studies had shortcomings in this paper. First, while the amount smaller in the data problem relatively, and inadequately the data processing in the early stage. In addition, the data does not involve missing and error, and it is relatively simple to handle.

We should focus on researching big data in the future, that improve the preprocessing of data, and pay attention to the processing of missing and missed data, which further enhance the prediction effect.
References

[1] Sun Yuxuan. Analysis of road traffic accidents based on data mining [D]. Beijing: Beijing Jiaotong University, 2014.

[2] Zhou Qian, Lu Huapu, Xu Wei. Traffic Accident Law and Its Model [J]. Journal of Traffic Engineering, 2006, 6 (4): 112-115.

[3] Zhang Weihua, Sun Hao, Mu Chaoxu. Traffic Safety Prediction Model and Simulation Based on Support Vector Machine [J]. Journal of System Simulation, 2009, 21 (19): 6266-6270.

[4] Wu C.H., Ho J.M., Lee D.T. Travel-time prediction with support vector regression [J]. IEEE Transactions on Intelligent Transportation Systems, 2004, 5 (4): 276-281.