Analysis and Forecast of Traffic Accident Big Data

Chen CHEN

Tunnel Traffic Engineering Design Office, Yunnan Province Transportation Planning and Design Institute, Kunming, China
522253113@qq.com

Abstract: Nowadays, as traffic accidents keep happening, traffic safety has become a major focus of contemporary social issues. Many factors account for traffic accidents, such as accident location, time period, driver's feelings, weather and other uncertain complex factors. As a result, the occurrence of traffic accidents is nonlinear, so it is necessary to explore the correlation between the data from many different aspects so as to avoid risks. By analyzing traffic data and graphics, R language shows how the data is related. After data preprocess, data selection by using R language Remap package remapB and remapH function, we get the locations of the accidents and the accident thermal chart, where you can find high- frequency accident locations. Besides, we employ decision tree, linear regression, random forest algorithm to model the data. According to the actual results, we can verify the correctness of the model and get the most accurate model and it can help us to predict this model with similar data in the future. The ultimate goal of data analysis is to choose the most accurate model after validating the model, analyzing the characteristics of the data and the relationship between the model and the data.

1 Introduction

1.1 Research Background

At present, China's national economy develops rapidly. Motor vehicle ownership, driving numbers, and road traffic flow continued to rise. Road traffic plays a more and more significant role in promoting economic and social development. The next traffic safety problem has become a key factor which can influence lives and property's safety of people, affecting and restricting the benefits of social and economic development. Road traffic safety is the result of the interaction of many non-deterministic factors such as people, cars, roads and the environment[1]. We need to analyze the road safety accident factors from many aspects and multi-level of the road. According to the characteristics such as accident occurrence, ambiguity and multi- factors, excavating such as the relationship, law, and features between all kinds of historical accident data. Our major cities facing a major problem is the safety of road traffic accidents. The analysis and forecast of large traffic data are conducive to planning transportation and improving the transport facilities significantly. Timely and accurate access to traffic accident data, analysis rapidly and establish the correct model become the primary requirements of large data traffic accident analysis.

The development and progress of road traffic have brought great convenience to human society, economic benefits and social prosperity. At the same time, the recurrence of traffic accidents has brought great disaster to human society. Traffic accidents have become a major public hazard in today's society. According to the traffic sector statistics, in 2011, road traffic accidents involved 210812 casualties, resulting in 62,387 people died and caused economic losses of 1078.73 million yuan directly. China is one of the countries which has the largest number of global traffic accidents.

In 2011, after the ban on drunk driving, the ownership of country car was 78 million, road traffic accidents were 210,812, the death toll as high as 62387. Comparing with Japan, the ownership of country car was more than 7,000, traffic accidents up to three times in China, while the death toll was only 4611 people. In the United States across the Atlantic, the ownership of the car was 285 million, the number of car accidents killed only 42,000 people[2]. China's road traffic development is facing the bottleneck period. There is still a huge gap between China and developed countries in traffic safety. Such as the development and enforcement of traffic regulations and the safety awareness of traffic drivers and the quality of the drivers themselves. We need to solve the problem of road traffic from a new perspective and direction in China.

Traffic accident with large data has the following characteristics:

Firstly, It has the amount of data and traffic is related to every citizen. Every day it will produce a lot of data.

Secondly, The data is incredibly complex, and it has a significant amount of data. The data attributes are complex, and garbage data occupy most of the need for a large number of data screening, deleting, and selecting[3]. To reduce the complexity of the data needs to obtain valid data efficiently.

Thirdly, for the data processing requirements quickly, high traffic accident data timeliness, and let the data more accurate, we need to process time efficiently.

The era of extensive data has come, intelligent traffic has become a major component of our daily life. More accurate traffic accident data and faster processing have emerged as a goal which data analyst pursue on.

1.2 Research Purposes

Large data has been into all aspects of traffic management deeply. It has a significant role especially in traffic accidents. In traffic trips, people are most concerned about traffic safety. So it has great significance to analyze the big data of traffic accidents. Traffic accidents have great importance on whether it is for people to travel recommendations, or the transport sector takes a reasonable and practical way to ease the traffic.

In this paper, the data is analyzed and processed, the invalid data is deleted, the dirty data is removed and the data of the traffic accident are analyzed, modeled, forecasted and verified. After the data is cleaned, the correctness of each model is verified and the data is displayed on the map. Find the law between the potential of data, through the adaptive method to choose a high degree of correctness model. Forecast the severity of traffic accident which occurred at a particular time and place to improve the effectiveness of the data.

1.3 Related research

Traffic big data as the emerging research areas, many scholars have a discussion on its origin and development status. Liu Youyuan (2009) discussed the data collection and analysis of traffic accidents in Australia. Meng Xiaofeng (2013) introduced the application and challenge of large data. Chen Mei (2012) discusses the use of large data in the transportation system. Tennessee patrols in the United States used IBM's predictable technology for traffic accidents, which introduced the system's CRASH version in 2014 (historical data collision reduction analysis). The state began deploying the system after installing the CRUSH from IBM and ESRI at the Memphis Police Department. After the police had deployed the system, it captured 70 violations within two hours, and 70 violations were usually the number of times that police trapped by the traditional way a week. By using the system, Tennessee road patrols provide the results show that the accuracy rate of up to 72% within the pilot range of 6 months.

Also, there is a P4S system developed in cooperation with the United States. The system is developed by SLD and basic road data and accident data and traffic environment data. It includes three levels of data analysis; one is the basic data analysis, the formation of the required frame; the other is developing the analysis tool by mining the primary data; the last one is application level, including traffic diagnosis, economic analysis, and traffic forecasting decisions. Using visual analysis to the traffic big data, and display the accident data on the map. After the data preprocessing (Filtering missing values, filtering data subsets, converting data types...), to compare the different data modeling algorithms. Using the vrious algorithms to build the model for the train set , and to pretic the results by the test set data. The prediction results of each model are compared with the actual data results, and the confusion matrix is given to compare the accuracy of each model with the kappa value (the fit between the observers).

2 Data Preprocessing

The data used in this paper is derived from the sample data of traffic accidents in Shanghai Public Security Bureau. The collection time of data is from July 2015 to April 2016 in the month, all of them is about the traffic accident. The total amount of data is more than 2000; the data contains four attributes, namely accident number, accident type, accident location and accident time. Analyzing the above data, among them the accident number is the character type, the type of error is text type, the accident location is text type and the time of the accident is time type. There is a missing value in the data, and the text data is not easy to handle.

2.1 Geographic Location is Converted to Latitude and Longitude

For the accident location attribute in the data, because of its value is the particular geographical area, it is necessary to transform the text type accident site into the numerical coordinate position[4]. Using Remap which is an interactive dynamic map data visualization tool based on the Echarts(a JS plugin made by the Baidu Company), proceed as follows:

Step 1, apply for Baidu map ak (that is, access to Baidu map API key), if you don't have Baidu account, you are required to register Baidu account access key.

Step 2, spelling http request url, pay attention to use the first step to apply for ak.

Step 3, after receiving the http request, you should return the data (json and xml format are supported).

Step 4, parse the data returned before (here using json format).

Step 5, convert and save the required data.

The following is an example of the conversion data, where Address is the address given in the original data, longitude latitude is the latitude and longitude after conversion:

Address: Shanghai Zhongshan South Road into the East Road, about 200 meters west

Longitude latitude：121.4616653 31.19660184

Address:50 meters west of Tongren Road, Beijing West Road, Shanghai

Longitude latitude：121.4565581 31.23211911.

2.2 Numerical Analysis

Preprocessing the data, analyzing the maximum and minimum values of the data, the mean, the distribution, and the overall impression of the data[5]. Including the handling and filling of missing values. In the analysis, it was found that there were six missing values in the type

attribute, and the missing values were interpolated. The analysis of the position can see the following minimum and maximum distribution.

Lowest:[Jiading] Ring Road Jihang Road East about 10 meters

[Jiading] Ring Road Jia Tang Road East about 30 meters

Edison Road into Zhang Heng Road NATO about 60 meters

An Bo Road Yingkou Road East about 30 meters Anhe Road New Land Road West About 2 meters

Highest: Zhuguang Road will be about 50 meters northbound Zhuo Road

Zhuguang Road into Songze Avenue NATO 100 meters

Zhuting Road into the Shanfang Road about 3000 meters north

Zhuting Road into the Yuanmen Road about 5 meters east

The west of the Zhu Fei Road connects with the Shangfei base secondary river bank road.

2.3 Consolidate Data

In this case, it is necessary to integrate the data that has been processed before. The text-type accident location in the original data is merged with the geographical coordinates of the site, that is latitude and longitude, it becomes a new data frame, which is used for the database. The values type of accident in the four essential attributes of the data is slight, general, significant, converting to numerical data according to the severity of the accident, the corresponding values are 1, 2, 3, and then the numerical type of a factor as a classification attribute. The time type in the data is converted to numeric data. Examples of processed data are as follows:

1 Shanghai City [Jiading] Ring Road, Jiaxing Road East about 10 meters 3.1e +15

1 2015/8/16 16:29 121.2516 31.39889.

3 Incident Data Visualization

3.1 Draw the Accident Map

The accident map is used to display the location of the accident distribution; It's the test data used to show the occurrence of traffic accidents in Shanghai. The data for this part is a randomly selected traffic accident location from July 2015 to April 2016.

Firstly, we get the map of Shanghai and then superimposed the traffic accident data into the map of Shanghai to get the following figure. Figure 1 is the map of Shanghai, Figure 2 is the accident distribution map.

Figure 1. The map of Shanghai

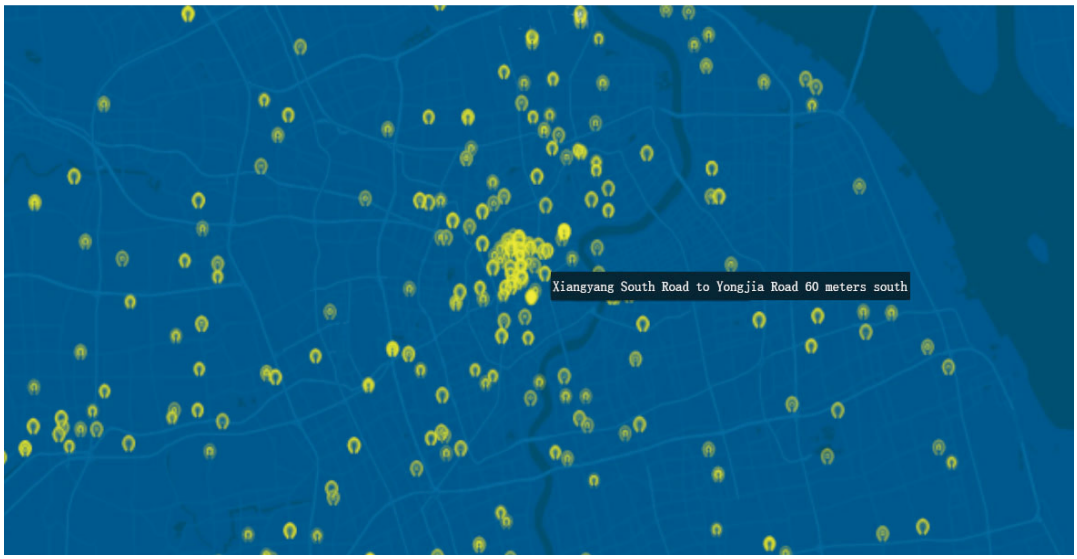

Figure 2. The Accident Map

3.2 Draw the Accident Heat Map

The accident heat map is used to show the traffic accident density in Shanghai. The data used is also derived from the previous test data that been randomly selected, It is the traffic accidents occurred in Shanghai from July 2015 to April 2016.

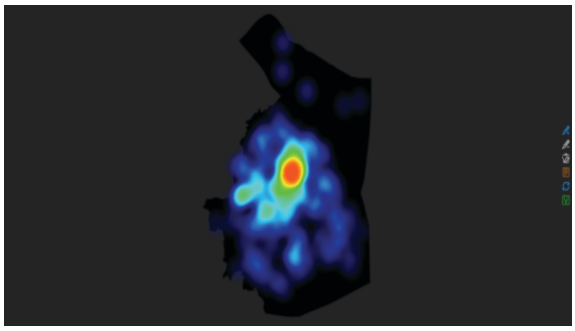

Figure 3. The Accident Heat Map

From the heat map, we can see that the distribution of accidents is centralized, combined with the numerical analysis of II.B , you can see the hot spots are mainly gathered in the accidental high incidence area around, which can be analyzed from data.

4 Modeling Prediction

This paper mainly uses the numerical types of the accident type as 1, 2 and 3 as the classification attributes to predict the different types of accident occurrence probability. The validation of the model uses K-fold cross validation. In the experiment, we use the 70% of the random data as the training data and the remaining 30% of the data as test data.

4.1 Fisher Linear Discriminant (FLD)

Fisher Linear Discriminant (FLD) is the classical algorithm in pattern recognition, which was used in pattern recognition and artificial intelligence field by Belhumeur in 1996[6]. The FLD algorithm project the high-dimensional pattern samples into the best discriminant vector space, to achieve the effect that extracting the classification information and compressing the feature space dimension. To guarantee the model samples have the largest outer-class distance and the minimum inner-class distance in the new subspace after projection, which requires that the pattern has the best separability in the current space. As an effective feature extraction method[7], it is possible to ensure that the next projection model has the smallest intraclass distance and the largest interclass distance in the new space, that is, the pattern has the best separability in the space.

4.2 Random Forest

Random forest is to build a forest as the random method. The forest consists of many decision trees, there is no association between each decision tree in the random forest. After getting the random forest, let the forest in each of the decision trees separately to make judgments when there is a new input sample into the forest[8]. Look at what kind of class this sample should belong to (for the classification algorithm), and then look at which kind of class was choice at most, to predict the sample is that class.

In the process of establishing each decision tree. There are two points to note: sampling and entire division. Firstly, it's the two random sampling process, the random forest needs to take a sample from the rows and columns of the input data. For row sampling, using the way that takes the back sampling methods, which means in the sampling set of samples, there are may be duplicate samples. Assuming that there are N input samples, then the sampling is also N samples. This situation makes it possible for each tree using not all of the samples when it's training, making it not easy to appear over-fitting relatively. And then column sampling, select m ($m << M$) from the M features[9]. The sampling data is used to establish a decision tree in an entirely split way. So that

the decision tree can not continue to divide or all the samples inside are pointing to the same one category. In general, many decision tree algorithms have an important step-pruning, But there's no need to do that because the two previous random sampling process ensured the randomness, so even if not pruning, it will not appear the over-fitting.

The experiment result of the random forest algorithm: Each decision tree is a skilled expert in a small field (because m features was chosen from M functions to train the each tree) So that there are many experts in various areas in the random forest. For a new question (new input data), data can be analyzed it from different aspects. By using the results of the random forest function prediction, there were 89 cases was predicted as 1 (the minor accident) and 4 cases were predicted as 2 (the general accident). There error predict result cases: there are one sample was predicted as the 1.0009203, but the actual result should be 1 (minor accident); there are 22 samples was predicted as the 2 (general accident), but the actual result should be 1 (minor accident); there are two samples was predicted as the 3 (major accident), but the actual result should be 1 (general accident); there are six samples was predicted as the 1 (general accident), but the actual result should be 2 (general accident).

4.3 Bagging Decision Tree

Decision Tree is based on the known of the probability of occurrence at the various situation, the probability that the expected value of the present value is greater than or equal to zero which obtained by constructing a decision tree to evaluate project risk, judge the feasibility of the decision analysis method. It is an intuitive graphical method of the probability analysis. Because this decision branch is painted like a tree branch, it is the called the decision tree.In the machine learning, the decision tree is a forecasting model, which represents a mapping between object attributes and object values. Entropy is the messy system degree, the ID3, C4.5 and C5.0 spanning tree algorithms using the entropy[10]. This measure is based on the concept of entropy in informatics theory.

A decision tree is a tree structure, which each internal node represents a test on an attribute value, each branch accounts for a test output, and each leaf node accounts for a category.

A decision tree is a very common classification method. It is a supervised learning method; the so-called supervised learning is given a bunch of samples, each sample has a set of attributes and a category, these categories are determined in advance, then we can get a classifier through the learning, the classifier will give the correct classification to the new samples. This machine learning method is called supervised learning.

Bagging:

Bagging algorithm- Creating a combined classification model for the learning strategy, where each model gives equal weight prediction.

Input:

D: d training tuples

K: the number of models in the combined classifier

A learning method (Decision tree algorithm, Backward propagation algorithm etl.)

Output: combined classifier-compound model MA

Method:

For i=1 to k do //Create K models;

Create a self-help sample Di by returning samples to D;

Using Di and the learning method to derive model Mi;

End for

Using the combined classifier to classify the tuple X: K models classified x, and return the majority of votes.

There are K training sets, and these training sets have the returning samples to create K models so that for each sample, every sample can be chosen[11]. The advantages of the bagging algorithm: 1. the accuracy rate of the combined model is significantly higher than any single classifier of the combined model. 2. For the larger noise, the performance will not be poor and has the robustness; 3.It's not easy to cause the over-fitting.

The correct result of the bagging algorithm: The result 1 (Minor accident) has 86 samples, the result 2(General accident) has four samples. The error result of the bagging algorithm: there are 8 samples was predicted as the 1 (Minor accident), but the actual result should be 2 (General accident); there are one sample that the real result should be 3 (Major accident); there are 22 samples was predicted as the 2(general accident), but the actual result should be 1(minor accident), there are 2 samples was predicted as the 3(Major accident), but the actual result should be 1(minor accident).

5 Experiment Analysis

By using LDA (linear judgment) function to build the model, and compare the predicted result with the actual result the result shows that the correct results have 95 cases. Even though the examination found, but the actual results are 1 (Minor accidents), but predicted to 1.00092 have one case, there are 26 cases of 2 (general), there are 2 cases of 3 (major) are wrong results.

The result of random forest, there are 89 cases was predicted to be 1 (Minor accident), there are 4 cases predicted to be 2 (General accident). The error predict result: there is one sample was predicted as the 1.0009203, but the actual result should be 1 (Minor accident); there are 22 samples was predicted as the 2 (General accident),but the actual result should be 1(Minor accident); there are 2 samples was predicted as the 3(Major accident), but the actual result should be 1(General accident); there are 6 samples was predicted as the 1(General accident), but the actual result should be 2(General accident).

The predicted result of tree bag function, the correct result: there are 86 cases of 1(Minor accident), 4 cases of 2 (General accident). The error predict result: there are 8 samples was predicted as the 1(Minor accident), but the actual result should be 2 (General accident), there are one sample that the actual result should be 3 (Major accident); there are 22 samples was predicted as the 2, but the actual result should be 1(Minor accident); there are two samples was predicted as the 3(Major accident), but the actual result should be 1(General accident).

In this paper, the predictive analysis of three algorithms is carried out to evaluate the accuracy of the model and the Kappa value (two observers fitting degree). From the minimum, quartile, mean and maximum values to compare the accuracy and Kappa values of the model(two observers fitting degree), the closer the accuracy of 1, the better the predicted result.

The Kappa value is between -1 to 1, but usually between 0 ~ 1. This is divided into five groups to represent different degree of consistency: 0.0 ~ 0.20 is the very low consistency, 0.21 ~ 0.40 is the general consistency, 0.41 ~ 0.60 is the moderate consistency , 0.61 ~ 0.80 is the high degree of consistency , 0.81 ~ 1 is the completely consistency.

To compare data from the accuracy and Kappa value, we can know the actual performance of the model, the graphic display is as follows.

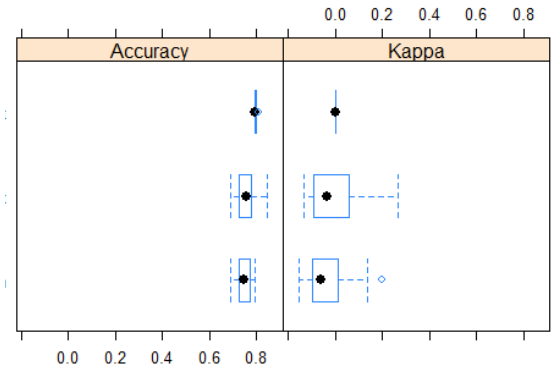

Figure 4. Compare results

6 Conclusion

From the above, we can find that the FLD algorithm has the best predict result from the accuracy aspect, the Bagging decision tree has the best predict result from the Kappa value, the random forest has the best predict result from the integrated effect. The experiment also reflects the characteristics of the three algorithms. Therefore, we can use the linear judgment analysis as the model, with the least error value when we the accuracy is more important; we can use the bagging decision algorithm when the appropriate degree is more important; we can use the random forests as the mode when the integrated effect is more important. The appropriate model should be determined by the different usage. Of course, this paper only uses three methods; we don't rule out the existence of other better models. For the current application model, linear judgment is the most suitable for traffic accident data prediction.

Acknowledgment

This work has been supported by the Open Foundation of Key Laboratory in Software Engineering of Yunnan Province under Grant NO. 2017SE204.

References

1. Qiao X M, An X U, Wei S. Developing tendency forecast of road traffic accident in China[J]. Journal of Changan University, 2004, 24(6):64-66.
2. Li G, Huang T Y, Yan H, et al. Grey residual error model of highway traffic accident forecast[J]. Journal of Traffic & Transportation Engineering, 2009, 9(5):88-93.
3. Qian W D, Zhi-Qiang L L, Prof. Road Traffic Accident Forecast Based on Gray-Markov Model[J]. China Safety Science Journal, 2008, 18(3):33-36.
4. Dong S H, Zhuo-Shen A P. Study on Road Traffic Accident Forecast Based on BP Neural Network[J]. China Safety Science Journal, 2010, 20(9):15-20.
5. Vuyst F D, Ricci V, Salvarani F. Nonlocal Second Order Vehicular Traffic Flow Models And Lagrange-Remap Finite Volumes[M]// Finite Volumes for Complex Applications VI Problems & Perspectives. Springer Berlin Heidelberg, 2011:781-789.
6. Quinlan J R. Induction on decision tree[J]. Machine Learning, 1986, 1(1):81-106.
7. Friedl M A, Brodley C E. Decision tree classification of land cover from remotely sensed data[J]. Remote Sensing of Environment, 1997, 61(3):399-409.
8. Quinlan J R. Bagging, Boosting, and C4.5[C]// Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference, AAAI 96, Iaai 96, Portland, Oregon, August 4-8. DBLP, 1996:725-730.
9. Wang H, Jiang Y, Wang H. Stock Return Prediction Based on Bagging-Decision Tree[C]// 2009 ieee international conference on the grey system and intelligent services. 2009:1575-1580.
10. Tu M C, Shin D, Shin D. A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms[C]// Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing. IEEE Computer Society, 2009:183-187.
11. Oza N C. Online bagging and boosting[C]// IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2006:2340-2345 Vol. 3.