ICIC Express Letters
Part B: Applications
Volume **10**, Number **7**, July **2019**

ICIC International ©2019 ISSN 2185-2766
pp. **619–625**

FATALITY RATE ANALYSIS OF MAJOR AND EXTRAORDINARY ROAD TRAFFIC ACCIDENTS BASED ON C4.5 DECISION TREE ALGORITHM

Tao Chen, Meidan Wu and Chu Zhang

Key Laboratory of Automotive Transportation Safety Techniques of Ministry of Transport
Chang'an University
Middle-section of Nan'er Huan Road, Xi'an 710064, P. R. China
chentao@chd.edu.cn

Received December 2018; accepted March 2019

Abstract. *The causation of road traffic accidents in China has been paid more attention recently. In order to better study the influencing factors of fatal road traffic accidents, this paper collected 250 cases of Major and Extraordinary Road Traffic Accidents (MERTA) with more than 10 deaths in China from 2005 to 2012. On the basis of the C4.5 decision tree algorithm, the five main influencing factors including purpose of vehicle, vehicle safety status, fault behavior, roadside safety facilities and driving age derived from the factor analysis theory were set as attribute conditions, and fatality rate was taken as a decision criterion to establish a fatality decision tree model. From the perspective of fatality rate, an in-depth analysis of the influencing factors of MERTA was illustrated. It was found that highway passenger vehicles and general freight vehicles had high accident rates, accounting for 49.6% and 18% respectively in 250 accidents. Drunk driving, fatigue driving, and over-speeding are high-risk acts of fault behavior. When the road is no roadside facilities or the driver's driving age is more than 10 years, the fatality rate is high. The results demonstrated that the most important factor affecting fatality rate was purpose of vehicle, followed by vehicle safety status, fault behavior, roadside protection facilities and driving age.*

Keywords: Major and extraordinary road traffic accidents, Decision tree model, C4.5 algorithm, Fatality rate

1. **Introduction.** In China, road traffic accidents have become a serious problem that threatens people's lives, property safety, and social sustainable development. The characteristics of road traffic accidents are as follows: the number of the injured and dead in one year is large, traffic accidents are seriously injured, and the fatality rate is high [1]. The situation of road traffic accidents is very worrying. Therefore, it is urgent to find out the main factors affecting the fatality rate.

Researchers abroad have done a lot of in-depth research on the analysis of the factors affecting the severity of traffic accidents and the construction of accident models, and have achieved fruitful research results. For example, Yau et al. [2] used logistic regression model to analyze the relationship between the severity of multi-vehicle traffic accidents and time, driver characteristics, vehicle types, road environment, and topographic conditions. The results showed that male drivers, accident time, speed limits and road types are significantly related to the severity of traffic accidents. Kraus et al. [3] used the Poisson model to analyze the influencing factors of roadside accident frequency. The results indicated that no left side shoulder had a significant impact on traffic accidents, and the central guardrail could reduce the occurrence of accidents on the left side of the road, and the geometric characteristics of the right side of the road affect the occurrence of accidents. Alam and Spainhour [4] found that human factors included use of alcohol,

DOI: 10.24507/icicelb.10.07.619

inattention, and high speed play a leading role in fatal traffic crashes. The survey results show that there are still weaknesses in current driving skills for young drivers. Domestic scholars have also accumulated certain achievements in the analysis of traffic accidents. For example, Sun et al. [5] of Beijing Jiaotong University put forward the concept of driver safety margin, established a hierarchical structure model of the causes of traffic accidents, and used Analytic Hierarchy Process (AHP) to determine the weight of causes. Taking account of the impact of people, vehicles, roads, and the environment on the accident, the main cause of the accident was objectively obtained. Xu et al. [6] from the Traffic Management Institute of the Ministry of Public Security studied how to use data mining methods such as correlation analysis, cluster analysis, and decision tree analysis to analyze traffic accident data in depth, which provide scientific decision-making basis for road traffic accident prevention and traffic safety management. However, most of the current existing road accident researches in China focus on how to reduce or avoid road traffic accidents, but rarely involve in the analysis of casualties caused by heavy traffic accidents. The loss of property caused by casualties is immeasurable.

In China, the fatality rate of a fatal traffic accident with more than 10 deaths has been fluctuating at 50%, while the death rate in China is around 20% [7]. Fatality rate reflects the degree of harm to people caused by road traffic accidents, which is an important indicator of people's insecurity in traffic accidents. The fatality rate caused by road traffic accidents is high and the consequences are serious, and it is urgently needed to give human attention in the society.

This paper mainly studies road traffic accidents with more than 10 deaths, which is defined as Major and Extraordinary Road Traffic Accident (MERTA). Combined with the C4.5 algorithm [8], 250 cases of China's MERTA from 2005 to 2012 were collected to analyze the influence of fatality rate. The five main accident impact factors derived from the factor analysis theory were regarded as an attributing condition to establish a fatality decision tree model. Through the comparative analysis and correlation analysis methods, the causes of the MERTA were quantitatively analyzed, and the main reasons for the fatalities were summarized.

2. Road Traffic Accident Data Collection.

The MERTA data forming the basis of this paper designs 30 information collection items and collects 250 cases of traffic accidents in China for 2005-2012 years. The collection items mainly include drivers, cars, roads, environment and other information. The overall situation is shown in Table 1.

Table 1. Statistical table of samples of MERTA

Year	Accidents number	Deaths number	Injured number	The sample number of deaths			The sample number of injured			Fatality rate
				10-20	21-30	> 30	10-20	21-30	> 30	
2005	47	807	705	34	11	2	29	13	5	53.37%
2006	38	558	463	31	6	1	29	6	3	54.65%
2007	26	389	449	21	5	0	19	3	4	46.42%
2008	29	476	504	22	6	1	19	5	5	48.57%
2009	24	329	345	22	2	0	18	3	3	48.81%
2010	34	461	432	32	1	1	23	8	3	51.62%
2011	27	455	404	21	4	2	18	7	2	52.97%
2012	25	361	367	23	1	1	17	4	4	49.59%
Total	250	3836	3669	206	36	8	172	49	29	51.11%

Using the "accident hierarchy tree" for fatality analysis can effectively excavate the importance of various types of influencing factors [9]. Accident level tree is a hierarchical structure composed of various factors related to traffic accidents. These factors have different classification characteristics, which can be divided into different levels. Of particular concern is that the same layer factors are subordinate to the upper levels factors and subdivided the next layer factors. This paper selects the five main factors derived from factor analysis theory as the second level factors, focusing on the influence of the second factors on fatality rate. The level tree of the accident factor is shown in Table 2.

Table 2. The list of accident factor hierarchy tree

The first level factors	The second level factors	The third level factors
Vehicle factors	Purpose of vehicle	Highway passenger transport; Dangerous goods cargo; Tourist passenger cargo; General freight; Rental passenger Carriage; Others
	Vehicle safety status	Normal; Steering failure; Brake failure; Tire blasting; Other mechanical failures
Accident factors	Fault behavior	Over-speed driving; Drunk driving; Fatigue driving; Illegal overloading; Others
Road factors	Roadside safety facilities	Unprotected facilities; Wave crash barrier; Anti-collision wall; Anti-collision pier
Personnel factors	Driving age	Within three years; Three to ten years; More than ten years

The fatality rate [10,11] defined in this paper is the number of deaths as a percentage of the total number of casualties in road traffic accidents. Because the data type of fatality rate is continuous, it is necessary to transform continuous data into discrete data when using decision tree algorithm. Therefore, before building the model, interval method is used to divide this attribute into two grades, as follows, low: $[0, 0.51)$; high: $[0.51, +)$, where the cut-off line of the fatality rate 0.51 is the average of the fatality rate of 250 accident cases.

3. **Fatality Decision Tree Model.**

3.1. **Algorithmic description.**

The decision tree is an instance-based induction learning algorithm. Classifying a large number of data by destination can find out the potential and valuable information for decision-making. The decision tree consists of a root node, several intermediate nodes and leaf nodes. The root node located at the top of the tree has the maximum sample size and the sizes of the other nodes decrease in turn. The following is a brief introduction of the process of recursively generating a decision tree from the root of the tree down one by one.

Input: A set of candidate attributes containing training samples.

Output: Decision tree diagram.

(1) Preprocessing the training attribute collection data, creating the root node of the tree, and determining the node attributes.

(2) Choosing the attribute with maximum information gain as test attribute in candidate set.

(3) Taking this attribute as the first node, the attribute value of the node as the branching point of the node, and clearing the attribute in the candidate set.

(4) Selecting the second node in the same way as above until the candidate set is empty.

The decision tree algorithm uses a tree structure to represent a data space, each branching node of which represents a classification. The classification process of decision tree will end until the test attribute is not found.

3.2. **The basic principle of C4.5 algorithm.** Considering that the data of road traffic accidents include qualitative and quantitative variables, the data type of the accident is very complicated and the fatality rate is continuous data; therefore, C4.5 algorithm is used to construct the decision tree.

The C4.5 algorithm [12] is created and improved by Qululan based on the ID3 algorithm. It can realize the discretization of continuous attributes and has the advantages of fast classification speed, high precision and easy understanding of classification rules. The most important is that this algorithm can find the most suitable splitting attribute for the decision tree. This paper proposes a road traffic accident classification algorithm based on the C4.5 decision tree to analyze the accident data. Not only can it determine the various attribute variables related to the cause of the accident, but also can realize the classification of the cause.

The principle of the C4.5 algorithm is as follows.

Let S be a set of n data samples $S = \{X_1, X_2, X_3, \ldots, X_n\}$, where each sample X_i can be described by the m-term attribute characteristics. Assuming that the attribute A_m has V different values, the sample set S is divided into V subsequences $(S_1, S_2, S_3, \ldots, S_v)$ according to different values of A_m.

(1) Calculating the expected values for classifying the sample date set: If the output of the total sample set S has p positive sets and n negative sets, then the average classification information of sample set S is as follows:

$$Info(p,n) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \tag{1}$$

(2) Calculating the amount of information divided by attribute A_m for each subsequence.

For the V subsequences of attribute A_m, if each subsequence S_i includes p_i positive examples and n_i counterexamples, the expected value $E(S_i)$ of the subsequence S_i divided according to the attribute A_m is expressed as follows.

$$E(s_i) = -\frac{p_i}{p_i+n_i}\log_2\frac{p_i}{p_i+n_i} - \frac{n_i}{p_i+n_i}\log_2\frac{n_i}{p_i+n_i} \tag{2}$$

(3) Calculating the information entropy classified by the attribute A_m and recording it as $Info(A_m)$.

$$Info(A_m) = -\sum_{i=1}^{v}\frac{p_i+n_i}{p+n}E(S_i) \tag{3}$$

(4) Calculating the amount of splitting information of attribute A_m for sample set S, denoted as $SplitInfo(A_m)$.

$$SplitInfo(A_m) = -\sum_{i=1}^{v}\frac{p_i+n_i}{p+n}\log_2\frac{p_i+n_i}{p+n} \tag{4}$$

(5) Calculating the information gain rate of attribute A_m and recording it as $GainRatio(A_m)$, which is divided according to the attribute A_m equal to the ratio of the information gain amount to the partition information amount, as follows.

$$GainRatio(A_m) = \frac{Info(p,n) - Info(A_m)}{SplitInfo(A_m)} \tag{5}$$

Calculate the information gain rate of the m attribute features corresponding to the set S and select the largest attribute factor A_m as the root node. Then, the different

classification values of the attribute A_m correspond to the V subsets (S_v) of S are recursively calculated and choose the attribute with the highest information gain rate as the root node of A_m. Branch nodes are continuously divided until the data types in all node subsets are the same, that is, leaf nodes are generated. The information gain rate represents the ratio of useful data in attribute factors. If the attribute gain rate corresponding to this attribute is larger, the proportion of useful data in the branch is larger, and this attribute can be used as the most suitable split point.

3.3. **Model establishment.** This paper uses the five main factors listed in the second level of Table 2 as attribute features to establish a fatality decision tree model. On this basis, an important grade that affects the fatality factors is discovered. First at all, it is necessary to discretize the continuous-valued attributes in the dataset, and the information gain ratio of each attribute characteristic is calculated and compared the sizes. Through calculation, it is concluded that the information gain rate of the purpose of vehicle is the greatest, so which is selected as the root node of the decision tree. When purpose of vehicle of the accident vehicle is road passenger transport, the C4.5 algorithm is used recursively to calculate the information amounts of the remaining attribute features in the case of selecting the road passenger transport. Similarly, the attribute with the highest information gain rate is selected as the sub node and so on until the leaf node is generated. The following is an output model of a decision tree using purpose of vehicle, vehicle safety status, fault behavior, roadside safety facilities and driving age as judging conditions and fatality rate as a criterion, as shown in Figure 1, where T represents 250 accident data sets.

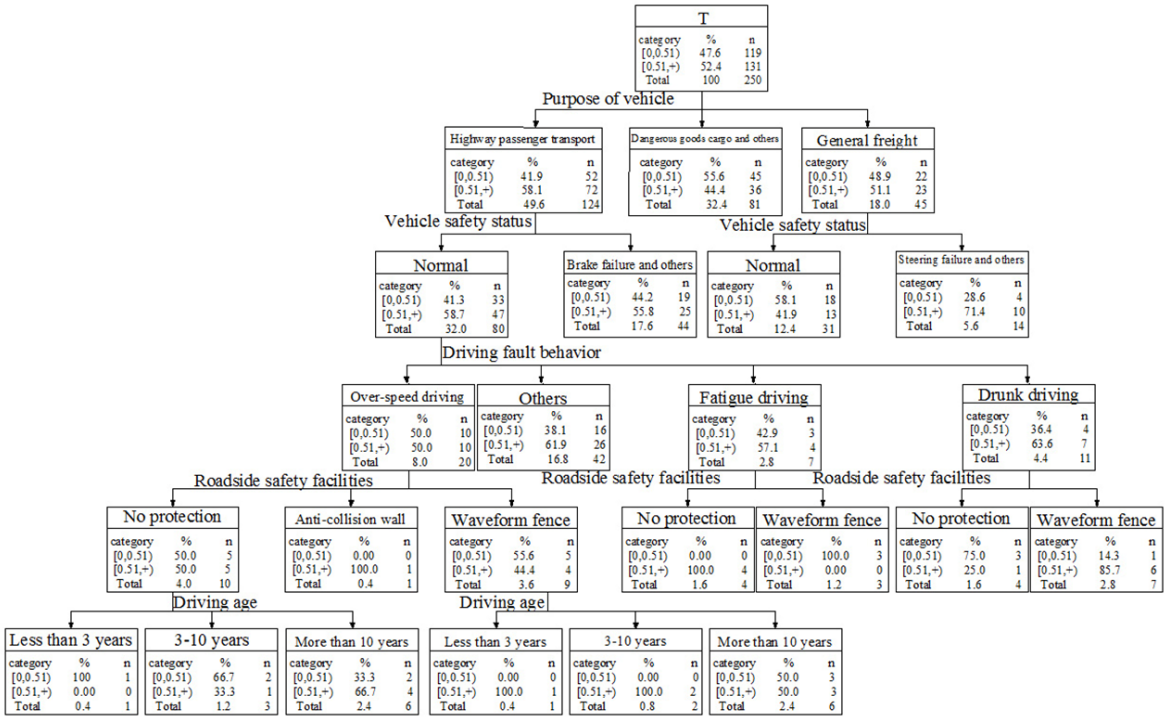

FIGURE 1. Fatality decision tree model

4. **Major and Extraordinary Road Traffic Accident Analysis.** Among the major influencing factors of people, vehicles, roads and environment, purpose of vehicle is located at the first layer of the decision tree, which is the most critical factor affecting lethality, followed by vehicle safety status, fault behavior, roadside safety facilities and driving age. According to Figure 1, an in-depth analysis of each layer of the decision tree is shown below.

(1) At the first layer of the decision tree, according to purpose of vehicle, the accident vehicles are divided into three categories: highway passenger transport, general cargo and other passenger transport. Among them, 124 accidents occurred on highway passenger transport, accounting for 49.6% of all accidents. Followed by general cargo, there were 45 accidents, accounting for 18.0% of all accidents, and the fatality rates of 23 accidents above 0.51.

(2) At the second layer of the decision tree, the node is the vehicle safety status. Among the 124 accidents in the selection of highway passenger transport, 80 accidents occurred during the vehicle driving normally, which accounted for 64.5% of all accidents, of which 47 accidents had a fatality rate of 0.51. In the general cargo vehicles, the proportion of accidents in the same conditions is the highest, accounting for 68.9%.

(3) At the third layer of the decision tree, the node is faulty behavior. Over-speeding, fatigue driving and drunk driving are the most frequent wrong driving behavior. The number of accidents of the three types respectively accounts for 25%, 8.75% and 13.8% of all accidents. In accidents caused by drunk driving, the proportion of accidents with fatalities above 0.51 was 63.3%.

(4) At the fourth layer of the decision tree, the nodes are roadside safety facilities. Over-speeding, fatigue driving, and drunk driving are all branched by the roadside safety facility. In the accidents of three wrong behaviors, accidents with unprotected facilities respectively accounted for 50%, 57.1%, and 36.4%. Second, the proportion of traffic accident on the roads of wave crash barriers was also high.

(5) At the fifth layer of the decision tree, the node is the driving age. Both unprotected facilities and wave crash barriers are branching out at the age of driving. In the unprotected accidents, accidents with a driving age of more than 10 years accounted for 60% of the total. In the case of the waveform crash barrier, the proportion was 66.9%. The fatal accident rate was about 50% when the driver's driving age was more than 10 years.

5. **Conclusion.** In this paper, 250 cases of MERTA were collected. A fatality rate decision tree model was established based on the C4.5 decision tree algorithm. The influencing factors of fatality rate of MERTA were analyzed in depth. The results show that among the main influencing factors of the accident, purpose of vehicle is the main factor that affects the fatality rate, followed by vehicle safety status, fault behavior, roadside protection facilities and driving age. When the vehicles using nature are road passenger transport or general cargo transportation, the proportion of accidents with a fatality rate of over 0.51 exceeds 50%. When these erroneous behaviors such as drunk driving, fatigue driving and over-speeding occur, the proportion of accidents with a fatality rate of over 0.51 reached 60%. When there are no protective facilities at the intersection and drivers are driving for more than 10 years, the fatality rate is high.

Acknowledgment. The authors acknowledge the support for this study provided by Science Foundation of Ministry of Transport of the People's Republic of China (20153198122 00), Chinese Universities Scientific Fund (300102228401, 300102228507), and Natural Science Foundation of Shaanxi Province, China (2016JM5013).

REFERENCES

[1] J. Y. Yang, Measures to reduce the fatality in traffic accidents, *Chinese Journal of Safety Science*, vol.12, no.5, pp.1-5, 2002.

[2] K. K. Yau, H. P. Lo and S. H. Fung, Multiple-vehicle traffic accidents in Hong Kong, *Accident Analysis and Prevention*, vol.38, no.6, pp.1157-1160, 2006.

[3] J. F. Kraus, C. L. Anderson and S. Arzemanian, Epidemiological aspects of fatal and severe injury urban freeway crashes, *Accid. Anal. Prev.*, vol.25, no.3, pp.229-239, 1993.

[4] B. Alam and L. Spainhour, Contributing factors for young at fault drivers in fatal traffic crashes in Florida, *Journal of Transportation Safety & Security*, vol.1, no.2, pp.152-168, 2009.

[5] P. Sun, R. Song and H. X. Wang, Analysis of the causes of traffic accidents on roads and countermeasures, *Safety and Environmental Engineering*, vol.14, no.2, pp.97-100, 2007.
[6] H. Y. Xu, Y. Q. Bao, H. L. Jiang et al., Research on data analysis and mining technology of road traffic accidents, *Journal of Chinese People's Public Security University: Natural Science Edition*, vol.14, no.4, pp.69-73, 2008.
[7] C. Zhang, *Analysis of China's Extremely Vicious Traffic Accidents*, Chang'an University, 2014.
[8] R. Naik, V. Kshirsagar, B. S. Sonawane et al., New strategy for detecting intrusion by using C4.5 algorithm, *Journal of Interventional Cardiology*, vol.22, no.3, pp.207-215, 2012.
[9] C. F. Zhang, Q. K. Bian and J. Chen, Fire risk analysis on university dormitory based on accident tree analysis and analytic hierarchy process, *China Safe Production Science and Technology*, vol.7, no.10, pp.100-105, 2011.
[10] T. Chen, C. Zhang and L. Wei, Regional freeway traffic safety evaluation for China based on driving volume and transportation volume, *Advances in Information Sciences & Service Sciences*, vol.4, no.15, pp.440-449, 2012.
[11] K. Mamady, B. Zou, S. Mafoule et al., Fatality from road traffic accident in Guinea: Aretrospective descriptive analysis, *Open Journal of Preventive Medicine*, vol.4, no.11, pp.809-821, 2014.
[12] M. Sudarma, The establishment of decision tree model in network traffic incident using C4.5 method, *Hepatology*, vol.24, no.3, pp.575-579, 2014.