Traffic accident analysis based on C4.5 algorithm in WEKA

Jiajia Li, *Jie* He*, *Ziyang* Liu, *Hao* Zhang, and *Chen* Zhang

School of Transportation, Southeast University, Nanjing 211189, Jiangsu, China

Abstract. At present, China is in a period of steady development of highways. At the same time, traffic safety issues are becoming increasingly serious. Data mining technology is an effective method for analysing traffic accidents. In-depth information mining of traffic accident data is conducive to accident prevention and traffic safety management. Based on the data of Wenli highway traffic accidents from 2006 to 2013, this study selected factors including time factor, linear factor and driver characteristics as research indicators, and established the decision tree using C4.5 algorithm in WEKA to explore the impact of various factors on the accident. According to the degree of contribution of each variable to the classification effect of the model, various modes affecting the type of the accident are obtained and the overall prediction accuracy is about 80%.

1 Introduction

As a passage connecting the cities, the highway undertakes huge traffic flow. Especially in China, the highway mileage is long, reaching 136,500 kilometers at the end of 2017 [1]. With the vigorous development of the highway industry, the traffic safety problem has become increasingly prominent. The daily personal injury and property damage caused by traffic accidents in China is very serious. The proportion of death tolls on highways to the number of road deaths in the country is on the rise, from less than 1% in 1994 to 10% in 2013 [2]. Therefore, the urgent problem that needs to be solved in the field of transportation today is to reduce the number of accidents and reduce accident losses. In order to achieve this goal, in addition to strengthening infrastructure construction and optimizing management measures, we also need to conduct research from the traffic accident itself to deeply understand the law of accidents.

Through the analysis of the causes of traffic accidents, targeted improvement measures can be proposed. The classification model can quantify the influence degree of each factor on the accident and clarify the influence mechanism of different accident types. It is beneficial for the analyst to grasp the key points.

In order to find out the main causes of accidents, this paper takes the statistical data of 8 years traffic accidents from K117 to K134 in Wenli highway as an example, and proposes a method based on C4.5 model to judge the main causes and accident types. The object of

* Corresponding author: hejie@seu.edu.cn

this study is to examine whether or not the C4.5 model can effectively identify the risk factors affecting types of accident.

2 Literature review

Numerous studies have focused on the factors affecting traffic accidents. After statistical analysis, the current mainstream method is to use data mining technology for analysis. From a methodological viewpoint, a wide variety of approaches have been employed to investigative traffic accident [3].

Sohn S Y et al. [4] used various algorithms to improve the accuracy of the classification of the severity of two types of road traffic accidents. His algorithms included classifier fusion based on Bayes and logic model; data integration fusion based on arc discharge and bagging, and clustering based on k-means algorithm. The research results show that the clustering-based classification algorithm is most suitable for the classification of road traffic accidents in Korea. Chang L Y et al. [5] established the CART classification tree model based on the traffic accident in Taipei in 2001 to investigate the relationship between accident severity and drivers, vehicle characteristics, environmental variables, and accident variables. The vehicle type was found to be most relevant to the severity of the collision. Li Y Z et al. [6] used the Apriori algorithm in the association rules to study the connection between traffic accident related factors. The accident data of Tianjin for 6 years was taken as the research object, and the obtained results were basically consistent with the inspection data of the traffic control department. Singh G et al. [7] examined the application of the M5 model tree and the conventional fixed/random effect negative binomial (FENB / RENB) regression model in the prediction of non-city sector accidents in the Haryana (India) highway. The results show that the two models perform quite well in terms of correlation coefficients and root mean square error values. The M5 model tree provides a simple linear equation that is easy to interpret. Lombardi D A [8] used the US Traffic Safety Administration's (NHTSA) summary of fatal traffic accident census data from 50 states in 2011-2014 to establish a multivariate regression and Poisson model to compare traffic accident-related factors between seniors and young drivers. Liu Z Q et al. [9] used NETICA software as a development platform to establish a Bayesian model to analyze the characteristics of highway traffic accidents in haze weather, and obtained the implicit relationship between the two.

In summary, the development of data mining in the field of traffic accidents presents a diversified trend, but there is a lack of exploration of the patterns of various types of accidents.

3 Methodology

The popularity of classification tree models stems from their widespread acceptability, ease of interpretation, and the provision of suitable estimation routines in the majority of popular statistical packages. In this study, the j48 algorithm in WEKA was used to explore the distribution of accident types. The j48 algorithm is also the C4.5 algorithm in the decision tree [10]. It is one of the most commonly used data mining techniques and is widely used in industrial and engineering fields [11][12]. A classification tree can be developed when the target variable is discrete. Because this study aims to simulate the distribution of accident types under various conditions in traffic accidents, and the results of distribution types are discrete, a classification tree has been developed.

The development of C4.5 decision tree model generally consists of the following three steps [13].

The first step is the growth of the branches. The growth of tree is also based on the information gain rate to segment the target variable. The calculation formula of the information gain rate is as follows:

$$gain(V)=\sum_{j}\frac{freg(C_j,T)}{\|T_i\|}lb(\frac{freg(C_j,T)}{\|T_i\|}-\sum_{j}\frac{\|T_i\|}{\|T\|}\sum_{j}\frac{\|C_{jv}\|}{\|T_i\|}lb\left(\frac{\|C_{jv}\|}{\|T_i\|}\right) \quad (1)$$

where V represents the set of attribute variables, T is divided into multiple subsets by V, and V has multiple different values. The number of examples of data set T is $\|T_i\|$. When $V=v_i$, the number of examples is $\|T_i\|$, the number of examples of C_j is $freg(C_j,T)$. When $V=v_i$, the number of examples is C_{jv}.

The second step is the processing of discrete variables. In the T set, $\{v_1,v_2,...v_n\}$ is the value of the continuous attribute A, and there are kinds of ways to divide A. The information gain rate of each division method is calculated, and the maximum information gain rate of A branch is assumed.

The third step is pruning. In order to prevent over-fitting of the model, the pre-selection pruning method is selected. In WEKA, it means to set the minimum number of instances of the branch. When the number of instances of the terminal node is too small, cut it off.

4 Data

According to the above method, the accident data of 2006-2013 collected from the K117 to K134 road sections of Wenli highway is taken as an example. Wenli highway is an important national trunk line in the central part of Zhejiang Province, connecting Lishui and Wenzhou. Wenli highway is known as the “Bridge and Tunnel Club” for its complexity of terrain. The 17km research section includes 5 tunnels and 1 bridge, and the geographical environment is complex. Figure 1 shows the study section scope.

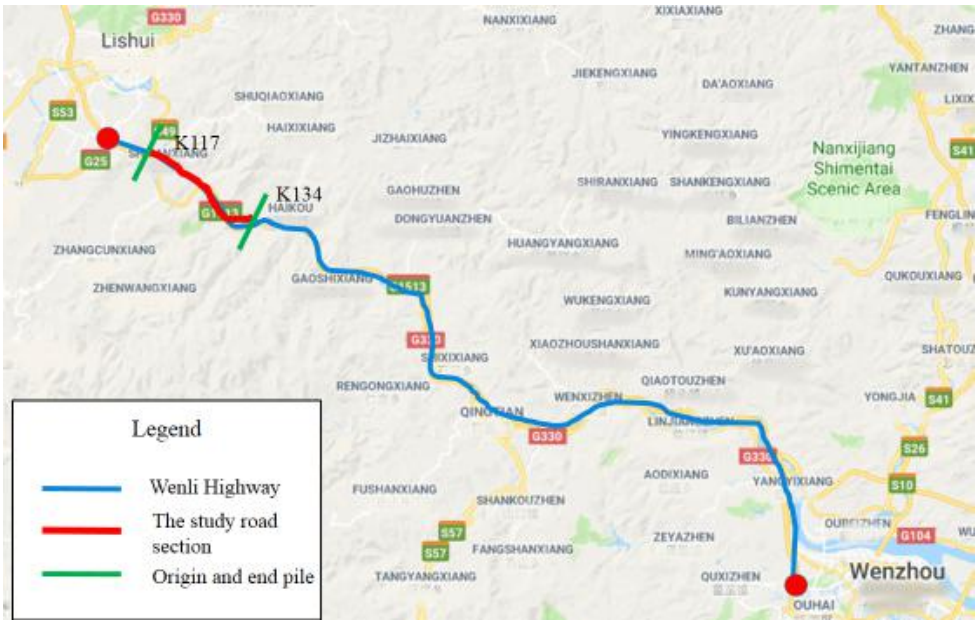

Fig. 1. Schematic map of study road section scope.

A total of 632 accidents were recorded on the road section. Table 1 shows the number of accidents occurred in each section. Each accident record includes time, place, vehicle type, record reason item, etc. The linear data, including the absolute value of the elevation difference per km and the radius of curvature, are obtained from designing documents.

Table 1. Number of accidents in each section of Wenli highway.

Start pile	End pile	Southeast line	Northwest line
K117+000	K118+000	6	10
K118+000	K119+000	5	19
K119+000	K120+000	4	45
K120+000	K121+000	22	28
K121+000	K122+000	8	23
K122+000	K123+000	12	15
K123+000	K124+000	33	36
K124+000	K125+000	29	29
K125+000	K126+000	18	33
K126+000	K127+000	7	23
K127+000	K128+000	8	16
K128+000	K129+000	7	17
K129+000	K130+000	3	9
K130+000	K131+000	17	13
K131+000	K132+000	11	12
K132+000	K133+000	21	15
K133+000	K134+000	22	23
K134+000	K135+000	10	23

The original data has the characteristics of clutter, incompleteness and ambiguity. After data cleansing, 586 accident records are finally obtained.

In order to build a C4.5 analysis model, the collected data needs to be divided into two subsets, one for training and one for testing. Normally, the training sample takes 2/3 and the test sample takes 1/3 [14]. Two subsets are selected by the method of generating a random number. A Mann-Whitney test shows that there was no significant difference in accident type between the two samples.

5 Calculation

This paper uses eight predictors to compare the target variables of the accident type, and try to find the important link between them. The 8 predictors include season, day of week, time of day, cause of accident record, vehicle type, terrain, radius of curvature and absolute difference in elevation. Table 2 gives the definition of variables.

Table 2. The definition of variables.

Variable		type	description	symbol
Input variable	Season	Discrete	spring	Spring
			summer	Summer
			autumn	Autumn

			winter	Winter
	Day of week	Discrete	workday	Workday
			weekend	Weekend
	Time of day	Discrete	6:00-24:00	N_night
			0:00-6:00	Night
	Cause of accident record	Discrete	speed too fast	Speeding
			loss of control	Uncontrolled
			rainy or snowy	Bad_weather
			poor car condition	Poor_condition
			fatigue driving	Fatigue driving
			Distance between vehicles is too small	Distance_small
			sprinkles from mountains	Sprinkles
			Puncture	Puncture
	Vehicle type	Discrete	cart or truck involved	Large
			only car	Car
	Terrain	Discrete	Ordinary terrain	Ord
			Bridge along the river	Bridge
			Tunnel entrance	Tunnel_entrance
			Inside the tunnel	Inside_tunnel
			Tunnel exit	Tunnel_exit
	Radius of curve	Continuous	500~12000	500~12000
	Absolute difference in elevation per km	Continuous	0~78.6	0~78.6
Output variable	Type of accident	Discrete	hit solid	Hit_solid
			rear end	Rear_end
			roll	Roll
			scratch	Scratch
			fire	Fire
			other	Other

The data is processed into a standard form and inputted into the WEKA software. The figure 2 shows the tree diagram obtained after the software running.

The tree has 13 terminal nodes, and collisions with solids and rear-end collisions are the main types of accidents. The first number in the parentheses of the terminal node is the correct classification, and the second number is the error classification. For example, the first branch of the model judges all types of accidents caused by rainy road slip as solid collisions, of which 44 are correct and 10 are wrong.

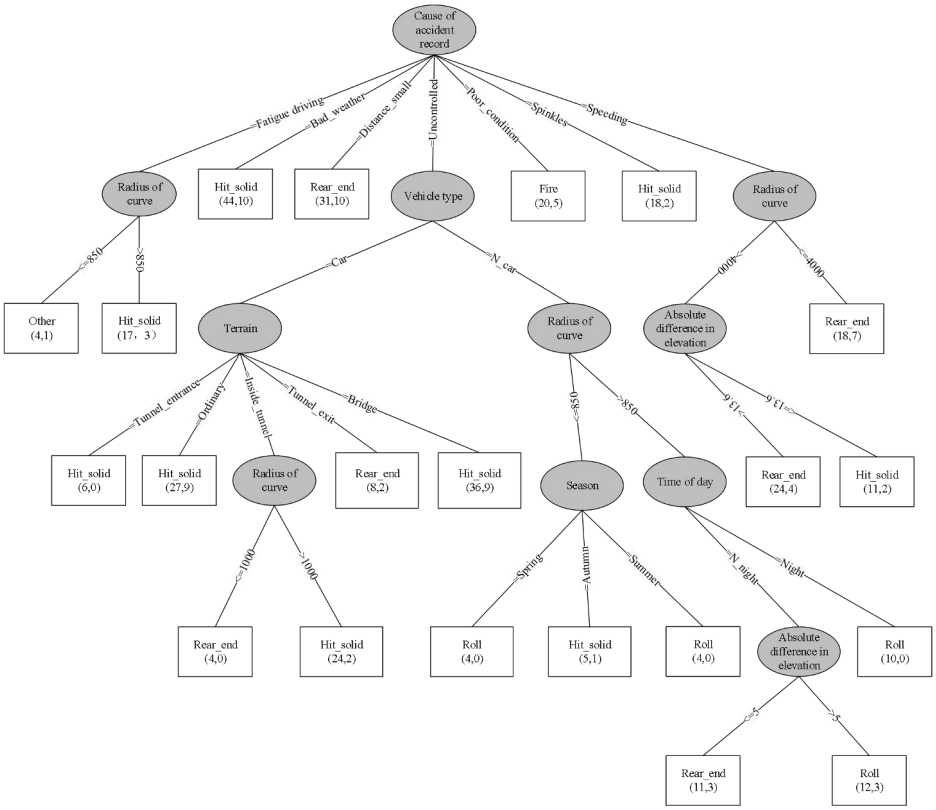

Fig. 2. The output of C4.5 tree.

The upper and lower order of the elliptical nodes represents the importance of each attribute. The causes of the accident record at the top are the primary factors, including speeding, loss of control, rainy or snowy, poor condition, fatigue driving, small distance between vehicles, sprinkles from mountains. It means this node contains the most important factors affecting the type of accident. Factors on the second level of the model are the radius of curve and whether there are only cars involved. The factors below the second layer have a weak influence on the classification effect.

The model can clarify the contribution of each element to the classification effect, In addition, it is possible to visually see from the tree diagram which type of accident is more likely to occur in some cases. Taking the rear-end accident as an example, the most common types of accidents in the following six cases are rear-end collision accidents.

(1) Cause of accident record: Small distance between vehicles;

(2) Cause of accident record: Speeding→ Radius of curve <=4000;

(3) Cause of accident record: Speeding→ Radius of curve > 4000 → Absolute value of elevation > 13.6;

(4) Cause of accident record: Loss of control → Cart or truck involved: No → Terrain: Tunnel exit;

(5) Cause of accident record: Loss of control → Cart or truck involved: No →Terrain: Inside the tunnel →Radius of curve: <=1000;

(6) Cause of accident record: Loss of control → Cart or truck involved: Yes →Radius of curve: > 850 → Time period: Daytime →Absolute value of elevation: <= 5.

This paper uses a simple correct rate to further describe the performance of the C4.5 model in determining the type of traffic accident. The overall prediction accuracy of the

training data is 81.34%, and the test data is 79.35%. The model has a higher accuracy rate in predicting both physical collision and rear-end collision types. However, in terms of other types of accidents, the prediction rate of this model is relatively unsatisfactory.

6 Discussion

The ultimate type of accident will always involve a complex interplay between a wide range of factors that are difficult to quantify. In this paper, the j48 algorithm in WEKA was used to construct the C4.5 model for analysis distribution of accident types. The model provides a good overall prediction of the training data and test data in this study, so the C4.5 model is a suitable method for analyzing the distribution of accident types. The C4.5 model can efficiently handle large data sets with a large number of explanatory variables and can produce useful results.

One of the advantages of the C4.5 model is to handle continuous variables, which is also its biggest advancement over the previous algorithm. In the C4.5 model, outliers are isolated into a node, so it do not cause splitting and may eventually be clipped. From a practical point of view, the results of the classification tree are displayed as graphical results, which will be easy for non-professionals to understand.

There are also some disadvantages in this model. First, due to the imperfection of statistical data, the ground slope of the accident occurred was replaced by the absolute value of the altitude per kilometer. In addition, the information about the driver was not collected, which reduced the accuracy of the model.

7 Conclusion

Using eight-year of vehicle accident data from Wenli highway, the model estimation results showed that the effects of the explanatory variables, involving road geometrics (radius of curve, absolute difference in elevation), vehicle type, terrain, and time factor (time of day, season).Various impact modes have been obtained for different types of accidents. This study shows that the C4.5 model is a suitable method for studying the type of traffic accidents.

Although some insights into the causal analysis of different accident types have been obtained, there are still some that need further study. At present, the information collection system in China is still at a relatively backward level, and more comprehensive and accurate information are needed in the future. And also, for a small sample of accident types, the model has a low correct rate. It can be considered to subsequently integrate a single type of accident for research.

Acknowledgments

The authors would like to thank National Natural Science Foundation of China (Grant No. 51778141), Transportation Department of Zhejiang Province (sponsored by project 2012H12), and Transportation Department of Henan Province (sponsored by project 2018G7). Their assistance is gratefully acknowledged.

References

1. Statistical yearbook of China[J]. 2017.

2. Gao Y, Dong X, Tian F. Analysis of current situation of highway traffic safety and management countermeasures[J]. China Safety Production Science and Technology. 2015, 11(10): 110-115.
3. Selvaraj S. Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques[C]// Iaeng-World Congress on Engineering and Computer Science. 2012.
4. Sohn S Y, Lee S H. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea[J]. Safety Science, 2003, 41(1):1-14.
5. Chang L Y, Wang H W. Analysis of traffic injury severity: an application of non-parametric classification tree techniques[J]. Accident Analysis & Prevention, 2006, 38(5):1019-1027.
6. Li Y Z, Zhang N, Liu F, et al. Application of Apriori Algorithm Based on Information Theory Optimization in Traffic Accident Analysis[J]. Information Systems Engineering. 2016(10), 80-84.
7. Singh G, Sachdeva S N, Pal M. M5 model tree based predictive modeling of road accidents on non-urban sections of highways in India[J]. Accident; analysis and prevention, 2016, 96:108.
8. Lombardi D A, Horrey W J, Courtney T K. Age-related differences in fatal intersection crashes in the United States[J]. Accident Analysis & Prevention, 2016, 99(Pt A):20.
9. Liu Z Q, Wang L, Zhang A H. Study on the Mechanism of Traffic Accidents in Wusongtian Highway Based on Bayesian Model[J]. Journal of Chongqing University of Technology. 2018(1), 43-49.
10. Markov Z, Russell I. An introduction to the WEKA data mining system[C]// Sigcse Conference on Innovation & Technology in Computer Science Education. ACM, 2006:367-368.
11. Yadav A K, Chandel S S. Solar energy potential assessment of western Himalayan Indian state of Himachal Pradesh using J48 algorithm of WEKA in ANN based prediction model[J]. Renewable Energy, 2015, 75:675-693.
12. Wang Hong-hai, 2013. Application of Classification Decision Tree in Analysis of Causes of Traffic Accidents[J]. Journal of Anhui University of Science and Technology, 2013, 27(06), 70-74.
13. Quinlan J R. C4.5: programs for machine learning[M]. Elsevier, 2014.
14. Breiman L. Classification and regression trees[M]. Routledge, 2017.