Driving risk assessment using near-crash database through data mining of tree-based model

Article *in* Accident; analysis and prevention · August 2015

DOI: 10.1016/j.aap.2015.07.007 · Source: PubMed

6 authors, including:

Jianqiang Wang
Tsinghua University
170 PUBLICATIONS **2,601** CITATIONS
SEE PROFILE

Yang Zheng
Harvard University
84 PUBLICATIONS **1,399** CITATIONS
SEE PROFILE

Chenfei Yu
Tsinghua University
6 PUBLICATIONS **121** CITATIONS
SEE PROFILE

Kenji Kodaka
6 PUBLICATIONS **120** CITATIONS
SEE PROFILE

Accident Analysis and Prevention 84 (2015) 54–64

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Driving risk assessment using near-crash database through data mining of tree-based model

Jianqiang Wang[a], Yang Zheng[a], Xiaofei Li[a], Chenfei Yu[a], Kenji Kodaka[b], Keqiang Li[a,*]

[a] State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing 10084, China
[b] Honda R&D Co. Ltd., Automobile R&D Center, Tochigi 321-3393, Japan

ARTICLE INFO

Article history:
Received 5 November 2014
Received in revised form 11 May 2015
Accepted 3 July 2015

Keywords:
Naturalistic driving study
Driving risk
Near-crash
Classification and regression tree (CART)
K-mean cluster

ABSTRACT

This paper considers a comprehensive naturalistic driving experiment to collect driving data under potential threats on actual Chinese roads. Using acquired real-world naturalistic driving data, a near-crash database is built, which contains vehicle status, potential crash objects, driving environment and road types, weather condition, and driver information and actions. The aims of this study are summarized into two aspects: (1) to cluster different driving-risk levels involved in near-crashes, and (2) to unveil the factors that greatly influence the driving-risk level. A novel method to quantify the driving-risk level of a near-crash scenario is proposed by clustering the braking process characteristics, namely maximum deceleration, average deceleration, and percentage reduction in vehicle kinetic energy. A classification and regression tree (CART) is employed to unveil the relationship among driving risk, driver/vehicle characteristics, and road environment. The results indicate that the velocity when braking, triggering factors, potential object type, and potential crash type exerted the greatest influence on the driving-risk levels in near-crashes.

1. Introduction

1.1. Background

In the past two decades, significant progress has been made in all aspects of vehicle safety systems, and experts from both academia and industry have conducted extensive research on vehicle safety (Young et al., 2014; Sepulcre et al., 2013; Takeda et al., 2011; Zheng et al., 2014). Efforts that aim to advance vehicle safety systems can mainly be divided into two areas (Jarašūniene and Jakubauskas, 2007): (1) active safety, which aims to avoid accidents and (2) passive safety, which helps reduce injuries in an accident. The active safety approach forecasts future driving states based on vehicle dynamics, infrastructure, and driver awareness (Wang et al., 2015), whereas the passive safety approach mainly focuses on enhancing vehicular safety systems such as seat belts, airbags and strong body structures (Jarašūniene and Jakubauskas, 2007). Although many encouraging achievements have been made, the number of road fatalities still remains unacceptably high, and traffic accidents are considered a major public health problem (DTM-China, 2010).

As the responsibility for traffic accidents involves the vehicles, drivers, and roadways, we must not only improve the safety performance of vehicles, but also better understand the factors that influence driving risk and identify the factors that result in accidents to make road transportation much safer. Many studies have attempted to better understand the factors that affect the probability and injury severity of crashes (Lord and Mannering, 2010). From a methodological standpoint, logit-based models are some of the most practical tools used for analyzing accident severity (Chen et al., 2012; Al-Ghamdi, 2002). Recently, non-parametric methods and data-mining techniques have been widely used to identify the factors associated with accident severity (Chang and Chen, 2005; Chang and Wang, 2006; Montella et al., 2011, 2012; Li et al., 2008; Harb et al., 2009). For example, Chang and Chen (2005) and Chang and Wang (2006) proposed a classification and regression tree (CART) model to establish the relationship among injury severity, driver/vehicle characteristics, and accident variables, indicating that vehicle type is a very important variable associated with cash severity. Li et al. (2008) evaluated the application of a support vector machine (SVM) model for predicting motor vehicle crashes, and showed that SVM models performed better than traditional negative binomial models. Montella et al. (2012) employed a decision tree and association rules to analyze accidents involving powered two-wheelers, and demonstrated that the curve alignment, rural areas, run-off-the-road crashes, night time, and rainy weather were

* Corresponding author.
E-mail address: likq@tsinghua.edu.cn (K. Li).

http://dx.doi.org/10.1016/j.aap.2015.07.007
0001-4575/

significantly associated with accident severity. These studies provided some insights into the factors that affect the likelihood of a vehicle accident. However, they were typically based on official traffic accident statistics, which have two major limitations: (1) lack of detailed driving data, and (2) difficult to collect and acquire (usually collected by traffic police agencies). Hence, the aforementioned studies usually do not consider the relationship between the accident severity and detailed driving data (*e.g.*, vehicle speed, acceleration, braking, and steering information).

Recent developments in vehicle instrumentation techniques have made monitoring naturalistic driving behavior and obtaining detailed driving data both technologically possible and economically feasible. For instance, NHTSA sponsored the project "100-Car Naturalistic Driving Study" which is a large-scale instrument-vehicle study to collect naturalistic driving data in the United States (Dingus et al., 2006). A series of technology tests of safety equipment was conducted in Michigan using the naturalistic driving technique (UMTRI and GMRDC, 2005). Takeda et al. (2011) reported a comprehensive project involving collecting large amounts of driving data on the actual road to study driver behavior and accident-causation-mechanism. With access to naturalistic driving data, traffic safety-related events could be observed and measured more precisely (Wu et al., 2014). Meanwhile, many researchers have proposed new methods and gained new insights into traffic safety (*e.g.*, Malta et al., 2009; Aoude et al., 2012; Guo et al., 2010; Jovanis et al., 2011; Jonasson and Rootzén, 2014). For instance, Malta et al. (2009) proposed a method to improve the understanding of driver behavior under potential threats using a large real-world driving database. Guo et al. (2010) assessed the factors associated with individual driver risk using naturalistic driving data. For naturalistic driving data, crash surrogates have received extensive research attention (see Guo et al., 2010; Wu and Jovanis, 2012, 2013; Moreno and García, 2013, for examples), because the number of crashes observed with naturalistic driving is typically small. Near-crash is frequently used as a surrogate measure for assessing the safety impact. For instance, Guo et al. (2010) employed two metrics, namely, precision and bias of risk estimation, to assess near-crashes, and indicated that using near-crashes as a crash surrogate could provide definite benefit when data about a sufficient number of crashes are not available. Recently, Wu and Jovanis (2013) proposed a multi-stage modeling framework to search through naturalistic driving data and extract near-crash events. All of these studies have demonstrated that naturalistic driving data could provide more controllable laboratory data as a useful supplement for traffic safety studies, and has the potential to further our understanding of crash causality, as well as improve road safety. Naturalistic driving data could not only provide more detailed driving exposure data, but also present the probability to identify more plausibly risky driving events and the associated factors.

1.2. Preview of the key results

This study focuses on the analysis of factors that influence driving risk using a naturalistic driving database. This database was obtained through designing a novel transcription protocol to code naturalistic driving data, which have two distinguishing features: (1) drivers drive in their normal states and (2) the instruments installed in vehicles can record drivers and road environments continuously during driving (Jovanis et al., 2011). The naturalistic database used herein contains only near-crash events because no actual crashes happened during the naturalistic experiments conducted on actual Chinese roads. Near-crashes refer to cases where drivers execute rapid evasive maneuvers (*i.e.*, emergency braking and/or steering operation) when facing a potential driving risk or a potential threat; in the absence of such an action, a real crash may occur. In the experiments, near-crash events in naturalistic driving were identified by detecting unusual vehicle kinematics using accelerometers and gyroscopic sensors installed in the experimental vehicle (Wu and Jovanis, 2013; Wu et al., 2014).

Recently, a few studies have focused on the assessment of risk in the driving environment, for example, individual driving risk (Guo and Fang, 2013) and momentary risk perception of a driving situation (Lu et al., 2012; Charlton et al., 2014). These studies employed indicators such as driver attributes and vehicle kinetic parameters to represent the risk level. Besides, critical braking and speed profiles were proposed to characterize the near-crashes in Moreno and García (2013) and Bagdadi (2013). In the present paper, we propose a novel method to quantify the driving-risk involved in a near-crash event. First, the driving-risk level is represented by the braking process characteristics, namely (1) maximum deceleration, (2) average deceleration, and (3) percentage reduction in vehicle kinetic energy. Then, the K-means cluster method is employed to classify near-crashes into different-risk levels based on the three aforementioned braking process features. Then, CART is employed for exploring the relationship among driving risk, driver/vehicle characteristics, and road environments. Identifying the factors associated with driving risk and further predicting high-risk driving scenarios will enable the adoption of proper safety countermeasures to reduce probable hazardous situations for high-risk groups, and thus improve overall driving comfort and safety. By analyzing driver characteristics, road conditions, and vehicle characteristics using the near-crash database, we obtained new insights into driving risk. The results indicate that the velocity when braking (V_BRA), triggering factors (T_FAC), potential object type (O_TYP), and potential crash type (P_CRA) had the greatest influence on the driving-risk level involved in near-crashes. These results can improve our understanding of the factors that affect driving risk, and help create polices and countermeasures to improve driving safety and comfort.

The remainder of this paper is organized as follows: Section 2 describes the near-crash database and presents some preparations, including experiment design, labeling protocol and driving-risk definition. The methodology employed in this study is presented in Section 3. Section 4 discusses the results, and some concluding remarks are given in Section 5.

2. Database and preparation

To build a firm foundation for the assessment of driving risk and enhancing driving safety, two components are essential: (1) real-driving data and (2) careful experimental design. Data collection is performed using naturalistic and low-intervention methods under actual traffic conditions. This section introduces the experimental equipment and experiment design, describes the near-crash database, and presents the definition and cluster analysis of driving risk.

2.1. Data-collection equipment and experiment design

2.1.1. Data-collection equipment

The naturalistic driving experiments were conducted using a Honda Crosstour, which was provided by Honda. The vehicle was equipped with instruments to collect driver, vehicular, and road data under real-world conditions. The data-collection system installed in the experimental vehicle included two driving recorders (DR) and four cameras (Fig. 1). The four cameras were used to record detailed video scenes including (1) forward view, (2) right-side forward view, (3) left-side forward view, and (4) driver's facial expression. One DR recorded data obtained by sensors, including GPS, brake signal, steering signal, three-axis

Fig. 1. Experimental vehicle and equipments.

acceleration information, and detailed video collected by the facial-expression and forward-view cameras. The other DR recorded the video collected by both the left-side and the right-side forward-view cameras to ensure convenient coding of the incidents.

In the present study, we focused on analyzing the driving risk of near-crash scenarios in the naturalistic driving experiments. Near-crash implies that the driver performs a rapid evasive maneuver (*i.e.*, emergency braking and/or steering operation), failing which a real crash may occur. In the experiments, near-crash events in naturalistic driving were identified by detecting unusual vehicle kinematics. For the experimental data collection under Chinese traffic environment, when the acceleration of the vehicle reached a threshold value (longitudinal: $-1.5\,m/s^2$, lateral: $-1\,m/s^2$), the data-collection system recorded the vehicle state (*i.e.*, speed, brake signal, steering signal, and three-axis acceleration), video sequence of the driver's expression, and video sequence of the events happening at the time. The recording time started approximately 10 s before the trigger point and lasted until 5 s after the trigger point. This means that each typical near-crash case has an approximately 15 s signal and video sequence. Fig. 2 shows a typical example of recorded driving signals. Note that the recorded video data should necessarily be reviewed manually to decide whether an event triggered by kinematic thresholds is actually safety-critical. If not, then such an event should be not defined as a near-crash and deleted from the dataset. For our experimental data, the recorded cases were checked mutually by Tsinghua and Honda.

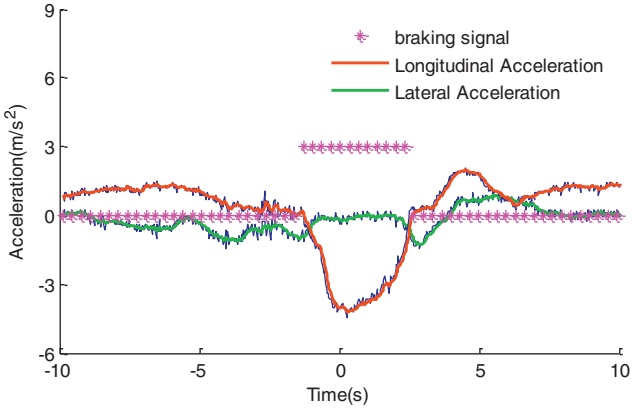

Fig. 2. Example of recorded driving signals for typical near-crash case.

Table 1
Schedule of entire experiment.

Time period	Morning	Afternoon	Night
Hours	140	220	50

Table 2
Road types in experiment.

Road type	1	2	3	4
Kilometers	1800	1210	4100	1650

1, highway; 2, city ring road; 3, inner-city road; 4, rural road.

2.1.2. Experiment design

The naturalistic driving route contained all road types, *i.e.*, inter-city highways (all structured road and usually low traffic volume), city ring road (mostly structured road and may have congestion), inner-city road (mixed traffic conditions and may be crowded with bicycles and motorcycles), and rural road (poor road structure and may be crowded with pedestrians). A total of 31 drivers, who signed the informed consent form, participated in the naturalistic driving experiments in their normal driving states. The experiment lasted 60 days, 6–7 h/day, resulting in naturalistic driving time and naturalistic driving range of approximately 400 h and over 8500 km, respectively. The schedule of the entire experiment plan is summarized in Table 1. Table 2 lists the naturalistic driving distance on the different road types considered herein.

Among the 31 drivers, 9 were female and 22 were male; all had regular driving licenses. The participants' average age was 43 years (ages ranging from 25 to 67 years) and they possessed a driving license for a mean period of 16 years (ranging from 3 to 48 years).

2.2. Labeling of near-crash database

Altogether, 912 near-crash events were recorded throughout the aforementioned 60-day naturalistic driving experiment. The distribution of these near-crashes by road type is summarized in Table 3. Deciding the protocol for labeling the multi-modal information is critical for properly associating near-crash driving situations with recorded driving state signals and videos. Following previous studies (Wu et al., 2014; Montella et al., 2012; Takeda et al., 2011) and considering actual traffic situations on Chinese roads, a novel data-transcription protocol that considers a comprehensive cross section of the factors that could affect the drivers and their responses is proposed in this paper. The proposed protocol comprises the following five major categories:

1) Vehicle status.
2) Potential crash objects.
3) Driving environment and road types.
4) Weather condition.
5) Driver information and driver actions.

The designed transcription protocol is comprehensive and contains important attributes that describe the conditions contributing to driving risk, providing potential for analyzing the relationship among driving risk, driver/vehicle characteristics, and road

Table 3
Near-crashes on different road types.

Road type	1	2	3	4
Number	39	246	489	138

1, highway; 2, city ring road; 3, inner-city road; 4, rural road.

Table 4
Definition of transcription protocol.

Variable	Code	Type	Description
Vehicle status			
Velocity when braking	V_BRA	Continuous	Vehicle speed when driver triggers braking signal or turn point of the acceleration signal (m/s)
Maximum deceleration	D_MAX	Continuous	Maximum deceleration during emergency braking (m/s^2)
Time interval of braking	T_IN	Continuous	Time interval between braking signal trigger and time point of maximum deceleration (s)
Velocity reduction	V_RED	Continuous	Vehicle-speed reduction from braking signal trigger to time point of maximum deceleration (m/s)
Vehicle status before braking	V_STA	Qualitative	1: Deceleration, 2: acceleration, 3: constant speed
Vehicle maneuver	V_MAN	Qualitative	1: Straight motion, 2: right turn, 3: left turn, 4: lane change, 5: others
Potential crash object			
Crash object type	O_TYP	Qualitative	1: Vehicle, 2: single-track vehicle (motorcycle and bicycle), 3: pedestrian, 4: others (*e.g.*, barrier block)
Potential crash type	P_CRA	Qualitative	1: Rear end, 2: conflict in intersection, 3: pedestrian conflict, 4: opposite driving conflict, 5: cut-in conflict, 6: others
Triggering factors	T_FAC	Qualitative	0: Sudden change of object status, 1: traffic light, 2: lane reduction, 3: lane change, 4: active braking, 5: others
Driving environment and road type			
Near-crash location	N_LOC	Qualitative	1: Intersection, 2: non-intersection
Road Condition	R_CON	Qualitative	1: Structure road, 2: normal road, 3: hybrid road, 4: rural road
Parking vehicle along road side	P_VEH	Qualitative	0: No, 1: yes
Safety barriers for opposing vehicles	B_OVE	Qualitative	0: No, 1: yes
Safety barriers for vehicles and pedestrians	B_VEH	Qualitative	0: No, 1: yes
Weather condition			
Weather	WEA	Qualitative	1: Sunny, 2:cloudy, 3: others
Light condition	L_CON	Qualitative	1: Daylight, 2: dusk
Driver information and actions			
Gender	GEN	Qualitative	1: Male, 2: female
Age	AGE	Continuous	Driver age (years). Further categorized into five groups, 1: 0–30, 2: 31–40, 3: 41–50, 4: 51–60, 5: >60
Time span with driving license	T_DIR	Continuous	Time period of possessing valid driving license (years)
Steering light	S_LIG	Qualitative	0: No, 1: yes
Vehicle horns	V_HON	Qualitative	0: No, 1: yes
Second Task	S_TASK	Qualitative	0: No, 1: talking, 2: others

environment. Graduate students with driving license served as volunteer taggers to manually label the recorded 912 near-crashes according to the designed transcription protocol. Finally, we developed the near-crash database. The transcription protocol is defined in Table 4. It should be noted that the specific definitions of each item in Table 4 are based on the actual characteristics of near-crash events and may have differences with the protocols for coding standard crash events, for example, that in Montella et al. (2013).

2.3. Definition and cluster of driving risks

The primary risk measure in vehicle safety evaluation is crash occurrence. Many studies have been conducted to identify the factors that significantly influence the injury severity of crashes using the logit-based model and some related data-mining techniques such as decision tree and SVM. However, research on naturalistic driving risk in the traffic and human-factor field has been limited.

In the present paper, driving risk is defined as a potential threat that may cause vehicle crashes or other accidents. Usually, the consequence of driving risk for a driver in his/her normal state is mainly reflected by rapid evasive maneuvers (*i.e.*, emergency braking and/or steering operation), which are employed by many studies on naturalistic driving to identify near-crashes, for example, Guo et al., 2010; Wu and Jovanis, 2013; Wu et al., 2014; Moreno and García (2013). In the naturalistic driving experiments conducted on Chinese roads, we found that nearly all the near-crashes had large longitudinal deceleration, implying the drivers tended to adopt the rapid baking maneuver to a avoid potential crash. Hence, the driving-risk level was represented by the braking process characteristics. Intuitively, the driving risk is higher if the braking maneuver is performed with greater urgency in a near-crash. By clustering braking process characteristics, this paper proposes a novel method to quantify the driving risk involved in a near-crash event. Fig. 3 shows the key points for defining a typical deceleration curve during braking. The following three features are adopted to represent the driving-risk level of a typical near-crash case:

1) Maximum deceleration during braking process $a_{\min}$.
2) Average deceleration $a_{average}$ from the braking trigger point t_0 to the point of maximum deceleration t_1.

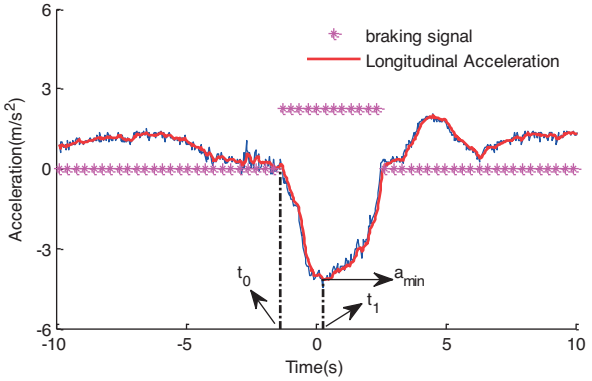

Fig. 3. Key features of driving-risk level.

3) Percentage reduction in vehicle kinetic energy η_E from t_0 to t_1.

The average deceleration $a_{average}$ is calculated as follows:

$$a_{average} = \frac{1}{t_1 - t_0}\int_{t_0}^{t_1} a(t)dt = \frac{1}{t_1 - t_0}\left[v(t_1) - v(t_0)\right], \quad (1)$$

where $v(t)$ and $a(t)$ denote the vehicle's velocity and acceleration, respectively.

The percentage reduction in vehicle kinetic energy η_E is calculated as follows:

$$\eta_E = \frac{(1/2)mv^2(t_0) - (1/2)mv^2(t_1)}{(1/2)mv^2(t_0)} = 1 - \left[\frac{v(t_1)}{v(t_0)}\right]^2, \quad (2)$$

where m denotes the vehicle mass.

Hence, the main criterion in evaluating the driving-risk level, namely, the braking process features, is obtained for each near-crash incident.

$$X = \left[a_{\min}, a_{average}, \eta_E\right]^T. \quad (3)$$

Cluster analysis is a valid approach for classifying driving risks involved in different near-crashes into different risk levels and has been used for assessing individual driver risk (Guo and Fang, 2013; Donmez et al., 2009). In this study, the K-means cluster method, which is used widely for cluster analysis in data mining, is employed to classify the driving risks involved in near-crashes into different risk groups based on the proposed feature X. Using a pre-determined number of clusters, the K-means cluster method partitions the observations into k clusters, where each observation belongs to a cluster whose mean is closest to its value (Kaufman and Rousseeuw, 2009). The K-means method minimizes the within-cluster sum of squares:

$$\underset{S}{\arg\min}\sum_{i=1}^{k}\sum_{X_j \in S_i}||X_j - {\mu_i}^2||, \quad (4)$$

where $X=[X_1, X_3, \ldots, X_n]$ is the set of observed data, which represents the feature $X_i = [a_{\min}, a_{average}, \eta_E]_i^T$ in the context of this paper; $S=[S_1, S_3, \ldots, S_n]$ represents the set of k clusters and μ_i denotes the mean point of cluster set S_i.

The driving-risk level in each near-crash case is placed in one of the following three groups: (1) low-risk group, (2) moderate-risk group, and (3) high-risk group. Near-crashes in the cluster with the highest maximum deceleration are placed in the high driving-risk group. The output of cluster analysis is shown in Fig. 4. Table 5 summarizes the statistical characteristics of the three driving-risk groups. The distribution of near-crashes belonging to the different risk groups follows a pyramid structure, which means that the high-risk group has the minimum events, whereas the low-risk group has the maximum number of events. We can observe that the maximum deceleration of the high-risk group is more than two times that of the low-risk group and the maximum deceleration of the moderate-risk group is much higher than that of the low-risk group, which make the cluster result reasonable.

3. Methodology

The aims of this study are as follows: (1) cluster near-crash cases by driving-risk level, and (2) assess the factors that influence the driving-risk level. Toward the first objective, feature extraction and K-means analysis were discussed in Section 2. For the second objective, classification and regression tree (CART) is employed to explore the relationship among driving risk, driver/vehicle characteristics, and road environment by using the obtained naturalistic driving database in Section 2. The details of the decision tree model and analysis techniques are discussed in this section.

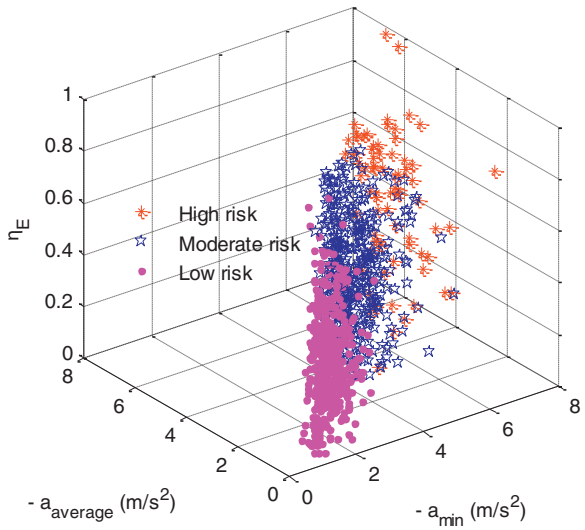

Fig. 4. Cluster result of driving risk.

3.1. Decision trees

Decision tree (DT) models are nonlinear and non-parametric data-mining tools, which can be used for supervised classification and regression problems. DTs are usually presented graphically as hierarchical structures, making them easy to understand. The main idea is to generate a DT using the known independent and target variables of a training dataset and then use the generated DT to predict the target variable of a new dataset. The DT structure can provide some insights into the relationship between the independent and target variables. Depending on the target variable, a classification tree (the target variable is discrete) or a regression tree (the target variable is continuous) is generated. This paper aims to model the driving risk involved in a near-crash event into discrete levels (low, moderate, and high), as discussed in Section 2. Hence, a classification tree is developed.

3.1.1. Decision tree structure

The main components of a DT include decision nodes, branches, and leaf nodes. Within a DT structure, each decision node represents a feature variable, and each branch stands for one of the states of this feature variable, which are based on the decision rules. The leaf node specifies the expected value of the target variable.

DTs are built recursively by partitioning a full dataset (noted by the root node) into a few small subsets using split criteria. The split criteria usually maximize the "purity" of the node dataset. Each subset is split until a pure state in the subset is reached such that its "purity" cannot be improved or the "purity" has reached a desired value. Pure subsets have no branches and no successor nodes. Thus, these subsets are called terminal or leaf nodes. When a new case or instance occurs, we can make a decision or prediction about the state of the case using its features and the tree structure. This procedure explains that the DT can be used to classify new cases, and the model structure can help us better understand the pattern behind the raw data.

3.1.2. Gini index and pruning

Different splitting indexes are available to show the main differences among DT-building procedures. One of the most famous splitting indexes is the Gini index, which is adopted in CART system. Given node dataset Y, the Gini index is calculated as follows:

$$Gini(Y) = 1 - \sum_i [p(Y=i)]^2, \quad (5)$$

Table 5
Characteristics of driving-risk groups.

Risk groups	Number of near-crash cases	Percentage	Mean of braking process features		
			a_{min} (m/s^2)	$a_{average}$ (m/s^2)	η_E
Low-risk	474	52.0%	−1.931	−1.027	30.9%
Moderate-risk	367	40.2%	−3.278	−1.717	56.6%
High-risk	71	7.8%	−5.385	−3.125	66.1%

where $p(Y=i)$ is the proportion of observations in node dataset Y belonging to class i. If all observations in one node belong to one class, the Gini index of that node is zero, which means that the node is pure and has reached a homogenous state.

The node-splitting criterion based on the Gini index aims to obtain the maximum decrease in the impurity of node dataset Y by finding the best partition x^* of observations, and then partition node dataset Y into two child node subsets Y_l and Y_r, as follows:

$$\max_{x\in X} \Delta Gini(Y,x)$$
$$\Delta Gini(Y,x) = Gini(Y) - p(Y_l)Gini(Y_l) - p(Y_r)Gini(Y_r) \quad (6)$$

where $\Delta Gini(Y,x)$ represents decrease in impurity, $x \in X$ denotes the set of splits generated by all features, Y_l and Y_r are, respectively, the left and right child nodes of node dataset Y; and $p(Y_l)$ and $p(Y_r)$ are the proportions of observations in node dataset Y belonging to the left and right child nodes, respectively.

Tree growing is arrested based on two criteria: (1) minimum decrease of impurity equals 0.001; and (2) maximum number of tree levels equals six. CART searches for the best split that maximizes (6). From this procedure, CART can be created recursively, which usually leads to saturation and overfitting of the training dataset. Saturated trees do not perform well when applied to a new case, which means that the tree structure overfits the information contained in the training data, including the useless noise information, and it cannot reveal the real pattern behind the data. Hence, the data are usually divided into two subsets: (1) learning (or training) set and (2) testing (or validation) set. The training set is used to construct the tree, and the testing set is used to validate the tree performance. The saturated tree should be pruned according to the cost-complexity algorithm that achieves a compromise between predictive accuracy and tree complexity. The main idea is to remove the branches and merge the nodes that contribute little to the predictive value of a tree. A more detailed description of the CART analysis and related applications can be found in Breiman et al. (1984). Analyses were performed using the SPSS software application.

3.2. Rule extraction

The CART structure can be transformed into decision rules of the 'IF–THEN' type to extract potentially useful information, which can be understood easily and intuitively by engineers and policymakers. Many researchers using DTs to analyze traffic accident severity have extracted useful rules for discovering behaviors that occur within a specified dataset (please see Montella et al., 2011, 2012; de Oña et al., 2013; Abellán et al., 2013 and the references therein).

The decision rules extracted from CART take a logic conditional structure '$X \rightarrow C$', where X denotes a set of statues of several attribute variables and C is the only statue of target variable, which is driving-risk level in our case. In CART, rules (IF–THEN structure) begin with the tree-root node, and each variable used in the splitting criterion for node partition generates the IF of the rules, which ends in leaf nodes with a THEN status. The THEN status is the status of leaf nodes that take the largest number of observations, which, in our case, is the driving-risk level.

3.3. Variable importance

One of the outputs of the CART technique is variable importance, which characterizes a variable's ability to influence the model. The relative importance of variable x_j is calculated as follows:

$$Vim(x_j) = \sum_{t=1}^{T} \frac{n_t}{N} \Delta Gini(Y_t, x_j), \quad (7)$$

where $Vim(x_j)$ denotes the relative importance of variable x_j; $\Delta Gini(Y_t, x_j)$ is the reduction in the Gini index obtained by splitting variable x_j at node t, according to (6); n_t is the total number of observations in dataset Y_t belonging to node t; N is the total number of observations; and T is the number of nodes in CART. The variable with the largest number according to (7) is regarded as the most important variable with respect to the others.

4. Results and discussion

4.1. Data distribution of driving-risk level

Nineteen predictor variables and one target variable (the driving-risk level) are used in the CART model to identify the important pattern that reflects the relationship among driving-risk level, driver/vehicle characteristics, and road environment. As can be inferred from Table 4, these 19 predictor variables include vehicle status (*e.g.*, vehicle maneuver), potential crash object (*e.g.*, crash-object type and triggering factors), driving environment and road types (*e.g.*, near-crash locations), weather condition (*e.g.*, weather and light condition), and driver information and driver actions (*e.g.*, driver gender and age).

Table 6 lists the information on driving-risk level in terms of the predictor variables, which indicates that traffic light in the fifth predictor variable T_FAC is an important factor affecting the driving-risk level because a relatively high proportion of near-crashes caused by sudden changes in traffic light status occurs in the moderate- and high-risk groups (55.3% and 35.0%, respectively). From the sixth predictor variable N_LOC, we find similar statistical results, where the proportions of near-crashes at intersections are relatively higher in the moderate- and high-risk groups (44.5% and 12.6%, respectively) than those away from intersection (38.0% and 5.2%, respectively). Other meaningful findings listed in Table 6 include finding that as the braking speed increases, the proportions of near-crash cases in the moderate- and high- risk groups increase. The proportions in the moderate- and high-risk groups are, respectively, 46.4% and 13.9%, when the speed at the braking point ranges from 10 to 20 m/s, whereas those when the speed at the braking point ranges from 0 to 10 m/s are, respectively, 34.7% and 2.8%, as shown under the 19th predictor variable V_BRA.

The aforementioned preliminary statistical results are consistent with the analysis result obtained from CART, which is presented in the next section.

4.2. CART analysis

For the CART model, the 912 near-crashes are randomly divided into two subsets – one for learning and the other for testing.

Table 6
Distribution of driving-risk levels by predictor variable.

Num	Variable code	Description	Count	Driving risk level			Num	Variable code	Description	Count	Driving risk level		
				LR 52.0%	MR 40.2%	HR 7.8%					LR 52.0%	MR 40.2%	HR 7.8%
1	V_STA	Deceleration	265	48.7%	43.8%	7.5%	9	B_OVE	No	372	58.9%	37.4%	3.8%
		Acceleration	531	55.4%	37.5%	7.2%			Yes	540	47.2%	42.2%	10.6%
		Constant speed	116	44.0%	44.8%	11.2%	10	B_VEH	No	537	55.9%	37.8%	6.3%
2	V_MAN	Straight motion	778	51.0%	40.4%	8.6%			Yes	375	46.4%	43.7%	9.9%
		Right turn	38	65.8%	31.6%	2.6%	11	WEA	Sunny	727	51.2%	41.3%	7.6%
		Left turn	41	65.9%	34.1%	0.0%			Cloudy	147	55.8%	34.7%	9.5%
		Lane change	46	45.7%	50.0%	4.3%			Others	38	52.6%	42.1%	5.3%
		Others	9	44.4%	44.4%	11.1%	12	L_CON	Daylight	796	52.3%	40.2%	7.5%
3	O_TYP	Vehicle	596	55.0%	40.4%	4.5%			Dusk	116	50.0%	40.5%	9.5%
		Single-track vehicle	98	72.4%	21.4%	6.1%	13	GEN	Male	661	51.0%	40.5%	8.5%
		Pedestrian	69	60.9%	37.7%	1.4%			Female	251	54.6%	39.4%	6.0%
		Others	149	22.1%	53.0%	24.8%	14	AGE	0–30	145	50.3%	41.4%	8.3%
4	P_CRA	Rear end	349	51.3%	45.0%	3.7%			31–40	291	54.0%	39.9%	6.2%
		Conflict during intersection	70	61.4%	32.9%	5.7%			41–50	232	48.7%	40.9%	10.3%
		Pedestrian conflict	65	60.0%	36.9%	3.1%			51–60	202	56.9%	35.6%	7.4%
		Opposite driving conflict	46	67.4%	28.3%	4.3%			>60	42	38.1%	57.1%	4.8%
		Cut-in conflict	191	63.4%	30.4%	6.3%	15	T_DIR	0–10	305	50.8%	40.0%	9.2%
		Others	191	63.4%	30.4%	6.3%			11–20	380	56.1%	37.1%	6.8%
5	T_FAC	Sudden change of object status	723	57.7%	37.6%	4.7%			21–30	157	46.5%	44.6%	8.9%
		Traffic light	103	9.7%	55.3%	35.0%			>30	70	47.1%	48.6%	4.3%
		Lane reduction	9	77.8%	22.2%	0.0%	16	S_LIG	No	784	51.3%	40.3%	8.4%
		Lane change	33	48.5%	48.5%	3.0%			Yes	128	56.3%	39.8%	3.9%
		Active Braking	26	57.7%	42.3%	0.0%	17	V_HON	No	859	51.1%	41.0%	7.9%
		Others	18	57.7%	42.3%	0.0%			Yes	53	66.0%	28.3%	5.7%
6	N_LOC	Intersection	317	42.9%	44.5%	12.6%	18	S_TASK	No	784	52.0%	40.4%	7.5%
		Non-intersection	595	56.8%	38.0%	5.2%			Talking	125	51.2%	39.2%	9.6%
7	R_CON	Structured road	285	46.0%	43.5%	10.5%			Others	3	66.7%	33.3%	0.0%
		Normal road	238	46.2%	43.3%	10.5%	19	V_BRA	(0, 10]	501	62.5%	34.7%	2.8%
		Hybrid road	251	62.5%	31.9%	5.6%			(10, 20]	388	39.7%	46.4%	13.9%
		Rural road	138	55.1%	43.5%	1.4%			(10, +∞]	23	30.4%	56.5%	13.0%
8	P_VEH	No	586	48.1%	42.8%	9.0%							
		Yes	326	58.9%	35.6%	5.5%							

Note: Num denotes the index of predictor variables, and LR, low-risk group; MR, moderate-risk group; HR, high-risk group.

Fig. 5 shows the classification tree generated by CART, where 70% of the entire observation set is applied for learning and the remaining observations (30%) are applied for testing, as in Montella et al. (2012) and de Oña et al. (2013). CART created a tree with 17 nodes and 9 terminal nodes. The decision rules extracted from CART are listed in Table 7. All probabilities of decision rules are higher than 52.0%, with 76.7% being the highest value (rule 1).

Table 7
Description of rules obtained from CART.

Node/rule	Rules CART: IF, . . .	THEN	Probability
3	IF (T_FAC = 2) AND (V_BRA <= 13.71)	MR	76.7%
4	IF (T_FAC = 2) AND (V_BRA > 13.71)	HR	63.2%
8	IF (T_FAC ≠ 2) AND (V_BRA <= 12.03) AND (O_TYP ≠ 1)	LR	76.0%
9	IF (T_FAC ≠ 2) AND (V_BRA > 12.03) AND (P_CRA ≠ 5)	MR	63.2%
10	IF (T_FAC ≠ 2) AND (V_BRA > 12.03) AND (P_CRA = 5)	LR	52.0%
12	IF (T_FAC ≠ 2) AND (V_BRA <= 12.03) AND (O_TYP = 1) AND (P_CRA = 4 OR P_CRA = 5 OR P_CRA = 6)	LR	69.2%
14	IF (T_FAC ≠ 2) AND (V_BRA < = 12.03) AND (O_TYP = 1) AND (P_CRA = 1 OR P_CRA = 2 OR P_CRA = 3) AND (V_RRA > 8.7)	MR	55.2%
15	IF (T_FAC ≠ 2) AND (V_BRA <= 12.03) AND (O_TYP = 1) AND (P_CRA = 1 OR P_CRA = 2 OR P_CRA = 3) AND (V_RRA <= 8.7) AND (AGE <= 1.5)	MR	57.1%
16	IF (T_FAC ≠ 2) AND (V_BRA < 12.03) AND (O_TYP = 1) AND (P_CRA = 1 OR P_CRA = 2 OR P_CRA = 3) AND (V_RRA <= 8.7) AND (AGE > 1.5)	LR	67.3%

Note: probability means percentage of observations in which the rule is accurate; LR, low-risk group; MR, moderate-risk group, HR, high-risk.

The root variable that generates the CART is T_FAC (see Fig. 5), indicating that the single best variable that classifies the driving-risk level is the triggering factor that leads to the braking maneuver. CART directs the triggering factor that involves traffic light to the left, forming node 1, and directs the remaining triggering factors to the right, forming node 2. For node 1 and depending on the braking speed (V_BRA), nodes 3 and 4 are obtained with different driving-risk levels. Near-crashes are high-risk (probability of 63.2%) if V_BRA is greater than 13.7 m/s, (rule 4) and moderate-risk (probability of 76.7%) if V_BRA is less than 13.7 m/s (rule 3). This result shows the direct relationship between moderate- and high-risk near-crashes and sudden changes in traffic lights with high vehicle speeds. This result is consistent with the statistical results presented in the previous section.

The rest of the rules are attributed to the triggering factors other than traffic lights (node 2). After this node, the CART is split according to V_BRA, and near-crashes with braking speeds of less than 12.03 m/s are sent to the left, forming node 5; the remaining cases are sent to the right, forming node 6. Based on the triggering factors and braking speed in node 6, nodes 9 and 10 are obtained depending on the potential crash type (P_CRA). If P_CRA denotes cut-in conflicts, the near-crashes are low-risk, with a probability of 52% (rule 10). However, if P_CRA is of another type, the near-crashes are of moderate risk with a probability of 63.2% (rule 9). In node 5, the CART continues to grow according to the potential object type (O_TYP). If O_TYP is not a vehicle after node 5, the near-crash case is low-risk with a probability of 76.0% (node 8 and rule 8). When the O_TYP is a vehicle (node 7), the CART is divided according to P_CRA. From this point in the CART structure, rule interpretation is difficult because multiple variables are involved in near-crashes. However, from the CART structure shown in Fig. 5, the following results are highlighted: if P_CRA is opposite driving conflict, cut-in conflict, or others, the near-crashes are low-risk with a probability of 69.2% (node 12 and rule 12). If P_CRA is rear end conflict, conflict during intersection, or jump out, CART is divided by V_BRA. At leaf node 14, if V_BRA is higher than 6.67 m/s, the near-crashes are of moderate risk (rule 14). For node 13, CART continues to split based on the driver age into leaf nodes 15 and 16. At leaf node 15, if the driver age is less than 30 years, the driving-risk level is moderate with a probability of 57.1% (rule 15). For the rest of the driver characteristics, leaf node 16 predicts the driving-risk level involved in near-crashes as low-risk with a probability of 67.3%. From this splitting process, the driving-risk level in near-crashes can be predicted by proceeding down the CART branches until a leaf node is reached.

To further understand the performance of CART, comparisons of model predictions between the observed and predicted risk levels for the learning and testing data are summarized in Table 8. The overall model prediction accuracy for the learning data is approximately 66% and that for the testing data is approximately 62%, which is within a reasonable range compared with the other studies on traffic accident severity in which classification methods were applied. For instance, Abdelwahab and Abdel-Aty (2001) used a neural network method and achieved accuracies of 65.6% and 60.4% in the training and testing phases, respectively. de Oña et al. (2013) obtained 55% and 54% accuracy when they applied DT using different algorithms (C4.5 and CART, respectively). The prediction performance of CART demonstrates that the CART structure can reflect the pattern hidden behind naturalistic data to some extent.

The main objective is to identify the risk factors that affect the driving-risk level using CART in conjunction with the near-crash database. The statistical results listed in Table 6, CART structure shown in Fig. 5, and rules listed in Table 7 present some clues and relationships. The next section discusses in depth the risk factors that affect driving-risk level.

4.3. Risk factors affecting driving risk

The variable importance obtained from CART is used to quantify the influence of potential risk factors on driving-risk level. Table 9 lists the normalized importance of these variables. Sixteen variables influencing the driving-risk level are detected, with values varying from 100% to 0.1%. It is observed that four variables, namely, (1) velocity when braking (V_BRA), (2) triggering factor (T_FAC), (3) potential object type (O_TYP) and (4) potential crash type (P_CRA), have the largest influence on the driving-risk level. Meanwhile, the other variables such as driver age (AGE), vehicle maneuver (V_MAN), second task (S_TASK), barriers for opposing traffic flow (B_OVE), and vehicles parked along the roadside (P_VEH) are considered to have relatively less effect in our study case.

4.3.1. Velocity when braking

As shown in Table 9, V_BRA is the most important variable affecting driving-risk level, which apparently does not agree with the results of previous traffic accident severity analyses. For example, lighting condition was considered to have the most important effect on the traffic accident severity (de Oña et al., 2013) and similar results were reported in Abdel-Aty (2003).

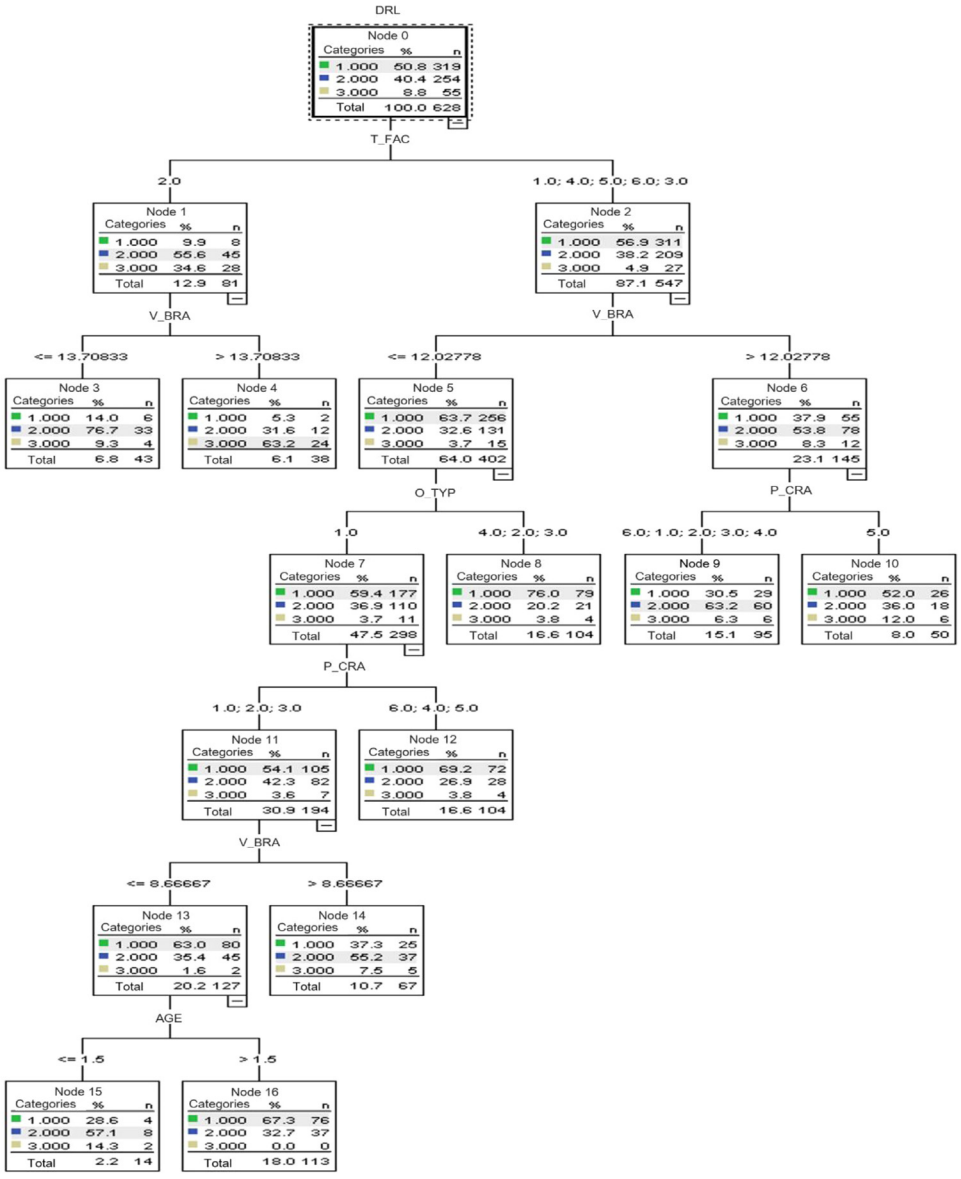

Fig. 5. Output of CART model.

Intuitively, the higher the vehicle speed, the higher is the kinetic energy of the lone-driver-vehicle system. If a potential threat is present or a sudden change in the object status occurs in the driving environment, the lone-driver-vehicle system becomes more unstable and risky, meaning that the driving-risk level involved in a near-crash case increases as the vehicle velocity increases. One direct explanation for these phenomena, in which vehicle velocity is usually not among the main factors that affect traffic accident severity, is that most traffic accident databases do not contain accurate speed information about traffic accidents (please see the database used in Chang and Chen, 2005; Chang and Wang, 2006; Li et al., 2008; Harb et al., 2009; Montella et al., 2012; Abellán et al., 2013).

On the other side, many studies have indicated that driving speed is an important factor for road safety (Elvik et al., 2004; Wallén and Åberg, 2008). Elvik et al. (2004) pointed out that speed not only affects the severity of a crash but is also related to the risk of being involved in a crash. From this perspective, our finding that

Table 8
Prediction result of CART model.

	Learning data ($N = 628$)			Testing data ($N = 284$)		
	Observed risk level	Predicted risk level	Correctly predicted	Observed risk level	Predicted risk level	Correctly predicted
Low-risk group	319	371	253 (79.3%)	155	167	113 (72.9%)
Moderate-risk group	254	219	138 (54.3%)	113	107	58 (51.3%)
High-risk group	55	38	24 (43.6%)	16	10	6 (37.5%)

The overall prediction accuracy is 66.1% for the learning data and 62.3% for the testing data.

Table 9
Importance of the predictor variable with CART (VIM).

Variables	Normalized importance
V_BRA	100.0%
T_FAC	96.7%
O_TYP	82.9%
P_CRA	75.9%
AGE	11.7%
V_MAN	9.0%
S_TASK	5.9%
B_OVE	5.8%
P_VEH	4.6%
N_LOC	2.7%
WEA	2.3%
V_HON	2.1%
B_VEH	1.1%
R_CON	0.5%
GEN	0.1%
V_STA	0.1%

V_BRA is the most important variable affecting driving-risk level agrees with those of previous studies in road safety research.

4.3.2. Triggering factors

Triggering factor (T_FAC) is the second most important variable with a normalized importance of 96.7% in the CART model (Table 9). Table 6 shows that traffic light in the fifth predictor variable T_FAC has a significant effect on the driving-risk level in near-crashes because a relatively high proportion of near-crashes caused by sudden changes in the traffic lights occurs in the moderate- or high-risk groups (55.3% and 35.0%, respectively). Rules 3 and 4 in Table 7 also support this finding. This result agrees with those of previous studies on vehicle crashes resulting from dilemma zones at signalized intersections (Rakha et al., 2007; Aoude et al., 2012).

4.3.3. Potential crash object and crash type

Crash object type (O_TYP) and potential crash type (P_CRA) have 82.9% and 75.9% normalized importance, respectively, in the CART model. Rules 9 and 10 in Table 7 demonstrate that at vehicle speeds greater than 12.03 m/s, the near-crashes caused by cut-in conflict (P_CRA is equal to 5, see definition in Table 4) are likely associated with the moderate-risk level, whereas the near-crashes caused by other factors are likely associated with the low-risk level group. Cut-in conflict usually occurs during lane change maneuvers. Lane change is an important factor that affects driving-risk level, and a lane change would lead to a collision if the maneuver is not proper (Pande and Abdel-Aty, 2006).

4.3.4. Other factors

Other factors such as driver age (AGE), vehicle maneuvers (V_MAN), second task (S_TASK), barriers for opposing vehicles (B_OVE), and vehicles parked along the roadside (P_VEH) have relatively small effects on the driving-risk level in our naturalistic driving experiment conducted on Chinese roads. It should be noted that this conclusion suits the driving environment in our naturalistic driving experiment. As the driving context changes, factors such as S_TASK, B_OVE, and P_VEH may have significant influences on road safety.

5. Conclusions

We recorded 912 near-crashes over the course of a 60-day naturalistic driving experiment involving 31 drivers. In this paper, a comprehensive transcription protocol containing important attributes that describe the conditions contributing to driving risk was first designed to analyze the relationship among driving risk, driver/vehicle characteristics, and road environment.

The main objectives of this study are as follows: (1) cluster driving-risk level and (2) unveil the factors that influence the driving-risk level. Toward the first objective, we proposed a novel method to quantify the driving-risk levels in near-crash cases by clustering the braking process characteristics, namely, (1) maximum deceleration, (2) average deceleration, and (3) percentage reduction in vehicle kinetic energy. K-means cluster analysis was applied to classify the near-crashes based on driving-risk level. Toward the second objective, CART is employed for unveiling the relationship among driving risk, driver/vehicle characteristics, and road environment using the near-crash database. CART provides an alternative and appropriate approach for analyzing driving-risk levels in near-crashes owing to its ability to identify hidden patterns in the data without pre-establishing a functional relationship among the variables.

Nine useful decision rules were obtained from the CART structure (Table 7). The overall model prediction accuracy for the learning data was approximately 66% and that for the testing data was approximately 62% (Table 8). These values are within the reasonable range compared with other studies on traffic accident severity. Furthermore, four variables, namely, (1) velocity when braking (V_BRA), (2) triggering factors (T_FAC), (3) potential object type (O_TYP) and (4) potential crash type (P_CRA), from CART were found to have the largest influences on the driving-risk level, which, to some extent, is in accordance with the results of some previous studies. These results validate the method proposed in this paper. It should be noted that there are some limitations of our current naturalistic driving experiment. First, it was only conducted in one city, *i.e.*, Beijing, and we carefully designed the experiment to include all the types of roads. Because of the actual road conditions in Beijing, however, there are few curves in our experimental routes. Hence, the influence of curve alignment could not be quantified in our current database. A few previous studies, for example, Montella et al. (2012) and Montella and Liana (2015), have pointed out that the curve alignment in road types was an important factor affecting road safety. Second, the time-duration of the current experiment was not very long (lasted for two months), and the weather conditions were sunny or cloudy for the most part. Intuitively, rainy weather would have a significant influence on the traffic safety, as pointed out in previous studies, for instance, Abellán et al. (2013). In our current database, the influence of weather conditions on the driving risk was not fully addressed. Despite such limitations, however, it should be pointed out that in this paper, the authors' proposed a novel method to quantify the driving risk in a near-crash event and to analyze the associated risk-factors. The proposed method can be extrapolated to specific studies on other datasets (*i.e.*, other infrastructure, roads, and countries).

Future research will consider individual driving risk because driving risk substantially varies among drivers and identifying factors associated with individual driver risk will further facilitate the identification of apt safety countermeasures (Guo and Fang, 2013). We would like to identify some factors such as age, gender, and driver characteristics that affect an individual driver risk by using the naturalistic driving data. Furthermore, O_TYP and P_CRA are also found to have important influences on the driving-risk level in near-crashes. One question worthy for further study is whether the factors that affect the driving-risk level remain the same when sub-datasets related to vehicles or pedestrians are the focus.

Acknowledgments

This collaborative research was supported by a joint project of Tsinghua and Honda. The authors would like to thank the National Natural Science Foundation of China (No. 51175290 and

No. 51475254) for its support. The authors would also like to thank those who participated in the driving experiments.

References

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. J. Saf. Res. 34 (5), 597–603.

Abdelwahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transp. Res. Rec.: J. Transp. Res. Board 1746 (1), 6–13.

Abellán, J., López, G., de OñA, J., 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Syst. Appl. 40 (15), 6047–6054.

Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accid. Anal. Prev. 34 (6), 729–741.

Aoude, G.S., Desaraju, V.R., Stephens, L.H., How, J.P., 2012. Driver behavior classification at intersections and validation on large naturalistic data set. Intell. Transp. Syst. IEEE Trans. 13 (2), 724–736.

Bagdadi, O., 2013. Assessing safety critical braking events in naturalistic driving studies. Transp. Res. F: Traffic Psychol. Behav. 16, 117–126.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press.

Chang, L.Y., Chen, W.C., 2005. Data mining of tree-based models to analyze freeway accident frequency. J. Saf. Res. 36 (4), 365–375.

Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid. Anal. Prev. 38 (5), 1019–1027.

Charlton, S.G., Starkey, N.J., Perrone, J.A., Isler, R.B., 2014. What's the risk? A comparison of actual and perceived driving risk. Transp. Res. F: Traffic Psychol. Behav. 25, 50–64.

Chen, H., Cao, L., Logan, D.B., 2012. Analysis of risk factors affecting the severity of intersection crashes by logistic regression. Traffic Inj. Prev. 13 (3), 300–307.

de Oña, J., López, G., Abellán, J., 2013. Extracting decision rules from police accident reports through decision trees. Accid. Anal. Prev. 50, 1151–1160.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J.D., Knipling, R.R., 2006. The 100-Car Naturalistic Driving Study Phase II-Results of the 100-Car Field Experiment (No. HS-810 593).

Donmez, B., Boyle, L.N., Lee, J.D., 2009. Differences in off-road glances: effects on young drivers' performance. J. Transp. Eng. 136 (5), 403–409.

DTM-China (Ministry of Public Security, Department of Traffic Management), 2010. Annual Report of Road Traffic Accidents Statistics in P.R. China. Scientific Research Institute of Traffic Management, Ministry of Public Security, Beijing (in Chinese).

Elvik, R., Christensen, P., Amundsen, A., 2004. Speed and road accidents. An evaluation of the Power Model. TØI Rep. 740, 2004.

Guo, F., Fang, Y., 2013. Individual driver risk assessment using naturalistic driving data. Accid. Anal. Prev. 61, 3–9.

Guo, F., Klauer, S.G., Hankey, J.M., Dingus, T.A., 2010. Near crashes as crash surrogate for naturalistic driving studies. Transp. Res. Rec.: J. Transp. Res. Board 2147 (1), 66–74.

Harb, R., Yan, X., Radwan, E., Su, X., 2009. Exploring precrash maneuvers using classification trees and random forests. Accid. Anal. Prev. 41 (1), 98–107.

Jarašūniene, A., Jakubauskas, G., 2007. Improvement of road safety using passive and active intelligent vehicle safety systems. Transport 22 (4), 284–289.

Jonasson, J.K., Rootzén, H., 2014. Internal validation of near-crashes in naturalistic driving studies: a continuous and multivariate approach. Accid. Anal. Prev. 62, 102–109.

Jovanis, P.P., Aguero-Valverde, J., Wu, K.F., Shankar, V., 2011. Analysis of naturalistic driving event data. Transp. Res. Rec.: J. Transp. Res. Board 2236 (1), 49–57.

Kaufman, L., Rousseeuw, P.J., 2009. Finding Groups in Data: An Introduction to Cluster Analysis, vol. 344. John Wiley & Sons.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accid. Anal. Prev. 40 (4), 1611–1618.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. A: Policy Pract. 44 (5), 291–305.

Lu, G., Cheng, B., Lin, Q., Wang, Y., 2012. Quantitative indicator of homeostatic risk perception in car following. Saf. Sci. 50 (9), 1898–1905.

Malta, L., Miyajima, C., Takeda, K., 2009. A study of driver behavior under potential threats in vehicle traffic. Intell. Transp. Syst. IEEE Trans. 10 (2), 201–210.

Montella, A.I., Liana, L., 2015. Safety performance functions incorporating design consistency variables. Accid. Prev. 74, 133–144.

Montella, A., Aria, M., Dambrosio, A., Mauriello, F., 2011. Data-mining techniques for exploratory analysis of pedestrian crashes. Transp. Res. Rec. 2237, 107–116.

Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F., 2012. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accid. Anal. Prev. 49, 58–72.

Montella, et al., 2013. Crash databases in Australasia, the European Union, and the United States: review and prospects for improvement. Transp. Res. Rec. 2386, 128–136.

Moreno, A.T., García, A., 2013. Use of speed profile as surrogate measure: effect of traffic calming devices on crosstown road safety performance. Accid. Anal. Prev. 61, 23–32.

Pande, A., Abdel-Aty, M., 2006. Assessment of freeway traffic parameters leading to lane-change related collisions. Accid. Anal. Prev. 38 (5), 936–948.

Rakha, H., El-Shawarby, I., Setti, J.R., 2007. Characterizing driver behavior on signalized intersection approaches at the onset of a yellow-phase trigger. Intell. Transp. Syst. IEEE Trans. 8 (4), 630–640.

Sepulcre, M., Gozalvez, J., Hernandez, J., 2013. Cooperative vehicle-to-vehicle active safety testing under challenging conditions. Transp. Res. C: Emerg. Technol. 26, 233–255.

Takeda, K., Hansen, J.H., Boyraz, P., Malta, L., Miyajima, C., Abut, H., 2011. International large-scale vehicle corpora for research on driver behavior on the road. Intell. Transp. Syst. IEEE Trans. 12 (4), 1609–1623.

UMTRI and GMRDC, 2005. Automotive Collision Avoidance System Field Operational Test Report: Methodology And Results. Final research report. NHTSA, US Department of Transportation.

Wallén, Warner H., Åberg, L., 2008. Drivers' beliefs about exceeding the speed limits. Transp. Res. F: Traffic Psychol. Behav. 11 (5), 376–389.

Wang, J.Q., Li, S.E., Zheng, Y., Lu, X.Y., 2015. Longitudinal collision mitigation via coordinated braking of multiple vehicles using model predictive control. Integr. Comput. Aided Eng. 22 (2), 171–185.

Wu, K.F., Jovanis, P.P., 2012. Crashes and crash-surrogate events: exploratory modeling with naturalistic driving data. Accid. Anal. Prev. 45, 507–516.

Wu, K.F., Jovanis, P.P., 2013. Screening naturalistic driving study data for safety-critical events. Transp. Res. Rec.: J. Transp. Res. Board 2386 (1), 137–146.

Wu, K.F., Aguero-Valverde, J., Jovanis, P.P., 2014. Using naturalistic driving data to explore the association between traffic safety-related events and crash risk at driver level. Accid. Anal. Prev. 47, 210–218.

Young, W., Sobhani, A., Lenné, M.G., Sarvi, M., 2014. Simulation of safety: a review of the state of the art in road safety simulation modelling. Accid. Anal. Prev. 66, 89–103.

Zheng, Y., Li, S., Wang, J., Wang, L., Li, K., 2014. Influence of information flow topology on closed-loop stability of vehicle platoon with rigid formation. In: 17th Intelligent Transportation System Conference, IEEE, pp. 2094–2100.