Accident Analysis and Prevention 129 (2019) 156–169

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Key feature selection and risk prediction for lane-changing behaviors based on vehicles' trajectory data

Tianyi Chen*, Xiupeng Shi, Yiik Diew Wong

School of Civil and Environmental Engineering, Nanyang Technological University, 639798, Singapore

ARTICLE INFO

Keywords:
Lane changing risk
Feature selection
Crash potential index
Resampling method
Random Forest

ABSTRACT

Risky lane-changing (LC) behavior of vehicles on the road has negative effects on traffic safety. This study presents a research framework for key feature selection and risk prediction of car's LC behavior on the highway based on vehicles' trajectory dataset. To the best of our knowledge, this is the first study that focuses on key feature selection and risk prediction for LC behavior on the highway. From the vehicles' trajectory dataset, we extract car's candidate features and apply fault tree analysis and k-Means clustering algorithm to determine the LC risk level based on the performance indicator of Crash Potential Index (CPI). Random Forest (RF) classifier is applied to select key features from car's candidate features and predict LC risk level. This study also proposes a method to evaluate the resampling methods to resample the LC risk dataset in terms of fitness performance and prediction performance. The cars' trajectory data collected from the Next Generation Simulation (NGSIM) dataset is used for framework development and verification. The sensitivity analysis of CPI indicates that the following cars in the original lane and target lane are respectively the safest and riskiest cars of the surrounding cars in an LC event. The results of resampling method evaluation show that SMOTETomek, which is less likely to be overfitting and has high prediction performance, is well suited for resampling the LC risk dataset on which RF classifier is trained. The results of key feature selection imply that the individual behaviors of the LC car and its surrounding cars in the original lane, the interactions between the LC car and its surrounding cars, and the interactions between the surrounding cars in the target lane (especially the interaction of the cars' accelerations) are of importance to the LC risk.

1. Introduction

Lane-keeping (LK) and lane-changing (LC) are essential maneuvers when driving on the road. LC occurs when drivers intend to obtain a more comfortable (or less congested) driving condition on an adjacent lane, or when merging or diverging across multi-lane traffic stream to reach the planned destination. LC has drawn increasing attention due to its negative effects on traffic safety (Zheng et al., 2013; Zheng, 2014). According to US road traffic accident statistics, LC and lane merge crashes account for 5 percent of police-reported crashes and 0.5 percent of fatalities in the US (Guo et al., 2010). The research on LC risk is of importance to the remediation of LC crashes, which can reduce property damage and life loss. However, research on LC is faced with several challenges. First, the rarity of crash data and inherent complexity make the study of LC more challenging than that of LK. Second, vehicle's LC behavior is not only influenced by preceding and following vehicles in the original lane but also preceding and following vehicles in target lane (Zheng et al., 2014). Third, several factors should be taken into consideration in an LC event, such as velocity and position of the LC vehicle and its surrounding vehicles, vehicles' characteristics, etc. (Hou et al., 2015).

This paper introduces a research framework for key feature selection and risk prediction of car's LC behavior on the highway based on vehicles' trajectory dataset. Key feature selection can provide a better understanding of the underlying process that the features contribute to the LC risk (Guyon and Elisseeff, 2003). The purpose of risk prediction is to predict LC risk level based on the features' performance. Considering the challenges mentioned above, we establish an LC risk dataset that comprises 1822 candidate features and the LC risk level of each LC event based on vehicles' trajectory dataset. For each LC event, we consider the trajectory data of the surrounding vehicles in both original and target lanes. The LC risk dataset is resampled by a resampling method to address the class imbalance problem before being trained by the machine learning classifier which is applied to select key features and predict LC risk level. Class imbalance problem occurs in a dataset when one or some of the classes contain many more samples

* Corresponding author.
E-mail addresses: TIANYI002@e.ntu.edu.sg (T. Chen), XSHI004@e.ntu.edu.sg (X. Shi), CYDWONG@ntu.edu.sg (Y.D. Wong).

https://doi.org/10.1016/j.aap.2019.05.017
Received 7 March 2019; Received in revised form 7 May 2019; Accepted 14 May 2019
Available online 28 May 2019

than others (Guo et al., 2016).

This paper is organized as follows. Section 2 introduces related works in the literature and discusses the research gaps that remain to be addressed. Section 3 describes the research framework and introduces the methods applied in each phase of the framework. Section 4 introduces the experimental verification of the framework and presents the findings. Section 5 summarizes the conclusions and future work.

2. Literature review

Recent research has studied LC from various perspectives. Some researchers focused on driver's behaviors in an LC event. Zheng et al. (2013) investigated the effect of LC in driver's behaviors and found that the transition behavior largely consists of a pre-insertion transition and a relaxation process. Li et al. (2016) proposed a novel method based on hidden Markov model and Bayesian filtering techniques to recognize driver's LC intention. Wang et al. (2018) identified LC warning threshold based on driver's perception characteristics when a rear vehicle is fast approaching. Some researchers focused on the mathematical models for LC trajectory, such as investigating LC trajectory on curved highway road (Guo et al., 2014), building LC trajectory models from driver's vision view (Zhou et al., 2017), and planning LC trajectory for automated vehicles (Yang et al., 2018). Some researchers applied game theoretic approach (Wang et al., 2015; Arbis and Dixit, 2019), microscopic simulators (Keyvan-Ekbatani et al., 2016), and fuzzy inference system (Balal et al., 2016) to explain the decision-making process of an LC maneuver, which is of significance to prevention of LC crash. Meanwhile, some researchers improved Advanced Driver-Assistance Systems (ADASs) (Butakov and Ioannou, 2015; Hou et al., 2015; Nilsson et al., 2016; Yan et al., 2016) and vehicle-to-vehicle communications (Luo et al., 2016; Chai et al., 2017) to assist with driver's decision-making process from a technical perspective.

Surrogate measures such as Time-to-Collision (TTC), Time Exposure Time-to-Collision (TET), Crash Potential Index (CPI), etc. (Mahmud et al., 2017) have been widely used to assess road traffic safety. For example, surrogate measures are widely applied in the estimation of vehicle rear-end crash risk. Oh and Kim (2010) proposed a method for estimating vehicle rear-end crash potential based on vehicle movements. Peng et al. (2017) used microscopic data and surrogate measures to assess the impact of reduced visibility on traffic crash risk and found that reduced visibility would significantly increase rear-end crash risk. Zhao and Lee (2018) demonstrated that CPI can effectively capture the rear-end collision risk and found that rear-end crash risk was lower for heavy vehicles than cars. Some researchers also focused on the performance evaluation and threshold determination of the surrogate measures in certain applications (Shi et al., 2018a; Rahman and Abdel-Aty, 2018). However, surrogate measures are mostly applied for safety analysis and crash prediction in LK scenario. Only a few researchers have attempted to improve surrogate measures and employ them in LC scenario. Park et al. (2018) proposed a new surrogate measure, Lane Change Risk Index (LCRI), to estimate crash risk in an LC event.

Machine learning methods have been widely applied in transportation studies, such as transportation mode recognition (Jahangiri and Rakha, 2015), congestion prediction in large-scale transportation network (Ma et al., 2015), traveling time prediction (Gal et al., 2017), and vehicle crash accident prediction (Shi et al., 2018b, 2019). Additionally, many researchers applied machine learning techniques, such as regression approach method (Schlechtriemen et al., 2015), Neural Network (Zheng et al., 2014; Izquierdo et al., 2017), Bayes classifier and decision trees (Hou et al., 2014), and Support Vector Machine (SVM) classifier (Izquierdo et al., 2017; Woo et al., 2017), to predict LC behavior and achieved satisfactory prediction accuracy. Most commonly-used machine learning methods assume that the sample size in each considered class is approximately similar. However, sample distribution is sometimes skewed since the representatives of some classes (e.g. risky behaviors, unusual patterns, etc.) tend to occur much less frequently in real-life, which can result in a severe class imbalance problem. The class imbalance problem poses difficulty for machine learning methods (Krawczyk, 2016; Guo et al., 2017).

From the literature scan, we find three research gaps that remain to be addressed. Firstly, few researchers have considered the longitudinal and lateral interactions between LC vehicle and its surrounding vehicles. Secondly, LC risk dataset always suffers from severe class imbalance problem, but few researchers have paid attention to this problem when analyzing LC risk. Thirdly, few studies have focused on key feature selection and risk prediction for LC behavior. To address these gaps, we firstly decompose vehicles' trajectories in longitudinal and lateral directions for vehicles in a multi-lane traffic stream and combine surrogate measures in a systemic approach to quantify LC risk. Then, we propose a method to assess the commonly-used resampling methods from which we identify the most suitable method for the LC risk dataset which suffers from class imbalance problem. Finally, key feature selection and risk prediction are conducted using the machine learning model trained with resampled LC risk dataset.

3. Methodology

3.1. Overall framework

The overall research framework is illustrated in Fig. 1. The framework mainly contains four phases namely, data preprocessing, candidate feature extraction, risk labeling, and key feature selection and risk prediction. The purpose of data preprocessing is to extract the LC trajectory data from vehicles' trajectory dataset. Candidate feature extraction aims to extract candidate features from the LC trajectory dataset. The phase of risk labeling contains two major steps of risk quantification and risk classification. Risk quantification is proposed to quantify LC risk, which involves surrogate measures and fault tree analysis based on the LC trajectory dataset. Risk classification applies unsupervised clustering method to classify quantified LC risk into several LC risk levels. In the final phase, supervised learning method is applied to select key features from candidate features and predict LC risk level based on resampled LC risk dataset. The four phases of the framework will be introduced in detail in Sections 3.2–3.5, respectively.

3.2. Data preprocessing

3.2.1. Definition of LC

In this study, we only extract and discuss LC events that involve five cars in each event, as shown in Fig. 2. Fig. 2(a) and (b) shows the scenarios of left-to-right LC and right-to-left LC, respectively. *sub* is the LC car and the location of it changes from the original lane to the target lane. *fol1* and *pre1* are the following car and preceding car of *sub* in the original lane, respectively. And *fol2* and *pre2* are the following car and preceding car of *sub* in the target lane.

3.2.2. Motion decomposition

In this study, we define the X, Y coordinates over the study road segment and decompose the car motions in the X and Y directions, which can clearly present the lateral and longitudinal interactions between LC car and its surrounding cars. This section explains how to decompose the velocity and acceleration of a car and the gap between two cars.

We assume that the velocity and acceleration of a vehicle are either in the same direction (i.e. when the acceleration is positive) or the opposite direction (i.e. when the acceleration is negative). Fig. 3 shows the locations of a car at the timestamps $t - \Delta t$, t, and $t + \Delta t$. The location of the car at the timestamp T is labeled by X_T and Y_T, which are the X, Y coordinates of the front-center of the car. V_T and A_T are the velocity and acceleration of the car at the timestamp T. V_{tX} and V_{tY} are the decomposed velocities of the car in the X and Y directions. V_{tY} and V_{tX} can be estimated as:

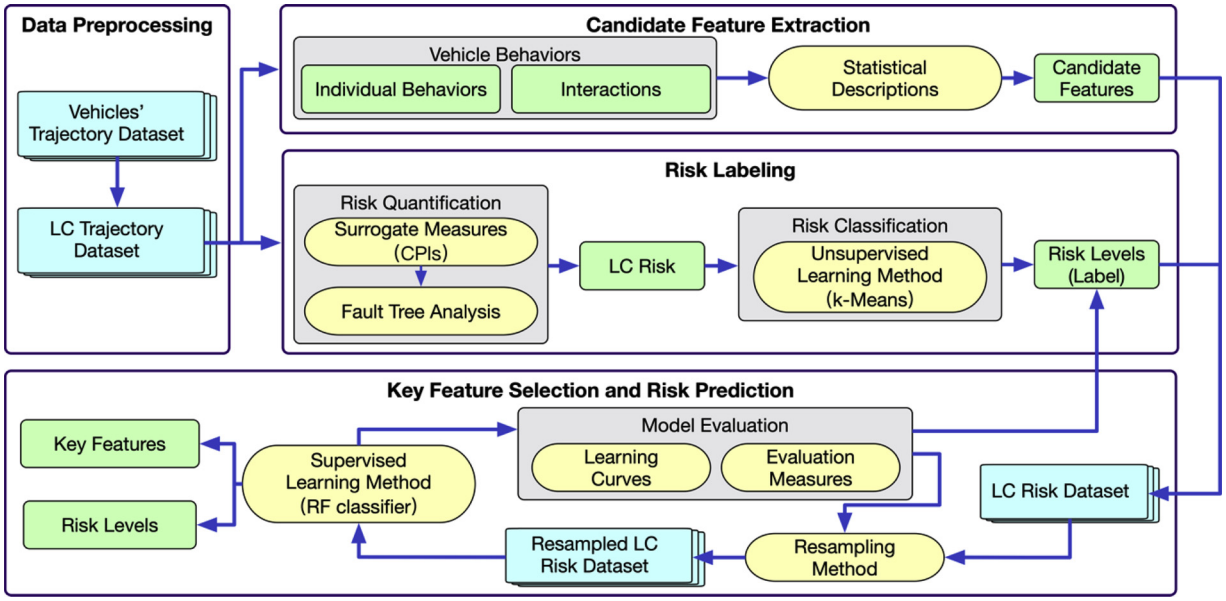


Fig. 1. Overall framework.

$$V_{tY} = \frac{Y_{t+\Delta t} - Y_{t-\Delta t}}{2\Delta t} \quad (1)$$

$$V_{tX} = \frac{X_{t+\Delta t} - X_{t-\Delta t}}{2\Delta t} \quad (2)$$

A_{tY} and A_{tX} can be estimated as:

$$A_{tY} = A_t \bullet \frac{V_{tY}}{V_t} \quad (3)$$

$$A_{tX} = A_t \bullet \frac{V_{tX}}{V_t} \quad (4)$$

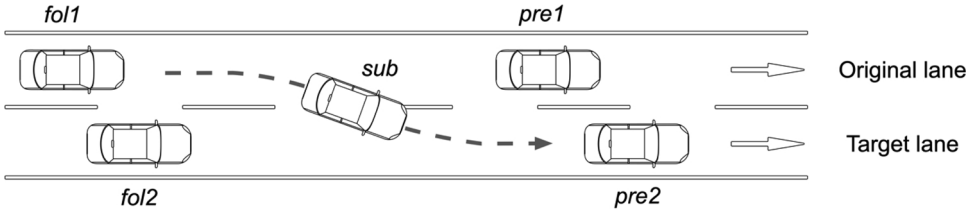
Fig. 4 shows the locations, widths, and lengths of the two neighboring cars, *Car1* and *Car2*. X_k and Y_k are the X, Y coordinates of the front-center of the *Car k*. L_k and W_k are the length and width of the *Car k*. In this study, we assume that the longitudinal centerline of a car is always parallel to the Y axis in an LC event on highway. Consequently, the gaps between *Car1* and *Car2* in the X and Y directions, G_X and G_Y, can be estimated as:

$$G_X = X_{car2} - X_{car1} - \frac{1}{2}(W_{car1} + X_{car2}) \quad (5)$$

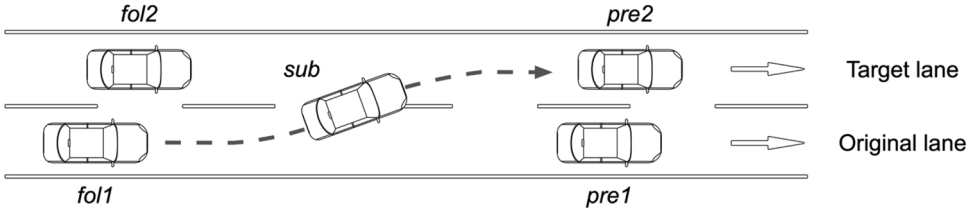
$$G_Y = Y_{car2} - Y_{car1} - L_{car2} \quad (6)$$

3.3. Candidate feature extraction

In this study, the statistical descriptions (e.g. mean, standard deviation, etc.) of the vehicle's behaviors of *sub*, *pre1*, *pre2*, *fol1*, and *fol2* in each LC event extracted from LC trajectory dataset are employed as candidate features. Vehicle's behaviors mainly comprise the motion behaviors of individual cars (e.g. location, velocity, acceleration, etc.) and the interactions between two cars (e.g. gap, velocity difference, acceleration difference, etc.). Vehicle's behaviors can also be categorized into general behaviors and LC behaviors. The general behaviors refer to the five cars' behaviors on the whole road segment, while the LC behaviors refer to the five cars' behaviors in an LC event. The candidate features also include the information of car length and car width.



(a) left-to-right LC



(b) right-to-left LC

Fig. 2. Illustration of LC scenarios.

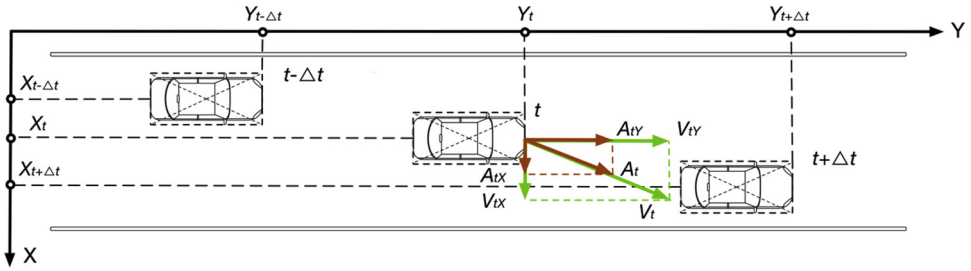


Fig. 3. Decomposition of car's velocity and acceleration.

We extract 1822 candidate features for each LC event, and key features shall be selected from these candidate features through machine learning. The candidate features are explained in detail in Table 1.

3.4. Risk labeling

3.4.1. Surrogate measures

a Deceleration Rate to Avoid a Crash (DRAC)

Deceleration Rate to Avoid a Crash (DRAC) is defined as minimum deceleration rate required by the following vehicle to match the velocity of the leading vehicle to avoid a crash (Cooper and Ferguson, 1976). In this study, DRAC is decomposed in the X and Y directions. As shown in Fig. 4, we assume that *Car1* would change lane from the upper one to the lower one and *Car2* would still travel in the same lane. The DRACs between *Car1* and *Car2* in the X and Y directions at the timestamp t, $DRAC_X^{car1\,\&\,car2}(t)$ and $DRAC_Y^{car1\,\&\,car2}(t)$, can be obtained as:

$$DRAC_X^{car1\,\&\,car2}(t) = \begin{cases} \dfrac{(V_X^{car1}(t) - V_X^{car2}(t))^2}{G_X}, & if\ V_X^{car1}(t) > V_X^{car2}(t) \\ 0, & otherwise \end{cases} \tag{7}$$

$$DRAC_Y^{car1\,\&\,car2}(t) = \begin{cases} \dfrac{(V_Y^{car1}(t) - V_Y^{car2}(t))^2}{G_Y}, & if\ V_Y^{car1}(t) > V_Y^{car2}(t) \\ 0, & otherwise \end{cases} \tag{8}$$

where $V_X^{car1}(t)$ and $V_Y^{car1}(t)$ are the velocities of *Car1* in the X and Y directions, $V_X^{car2}(t)$ and $V_Y^{car2}(t)$ are the velocities of *Car2* in the X and Y directions.

- Maximum Available Deceleration Rate (MADR)

Maximum Available Deceleration Rate (MADR) is defined as the maximum threshold of DRAC (Cunto and Saccomanno, 2008). *Car1* is deemed to 'hit' *Car2* in the X or Y direction if $DRAC > MADR$ in that direction. The probability that *Car1* 'hits' *Car2* in the X and Y direction at the timestamp t, $P_X^n(t)$ and $P_Y^n(t)$, are assumed to follow Bernoulli distribution, which is:

$$P_X^{car1\,\&\,car2}(t) = \begin{cases} 1, & if\ DRAC_X^{car1\,\&\,car2}(t) > MADR_X^{car1\,\&\,car2}(t) \\ 0, & Otherwise \end{cases} \tag{9}$$

$$P_Y^{car1\,\&\,car2}(t) = \begin{cases} 1, & if\ DRAC_y^{car1\,\&\,car2}(t) > MADR_Y^{car1\,\&\,car2}(t) \\ 0, & Otherwise \end{cases} \tag{10}$$

where $MADR_X^{car1\,\&\,car2}(t)$ and $MADR_Y^{car1\,\&\,car2}(t)$ are the MADRs between *Car1* and *Car2* decomposed in the X and Y directions.

However, the values of MADR are determined differently by researchers and organizations. Cunto and Saccomanno (2008) assumed that the MADR follows a truncated normal distribution with average equal to 8.45 m/s^2, standard deviation equal to 1.4 m/s^2, upper limit equal to 12.68 m/s^2, and lower limit equal to 4.23 m/s^2. The Institution of Transportation Engineers recommends that the maximum deceleration of car is 3.0 m/s^2 (Maurya and Bokare, 2012) and AASHTO (2001) recommends that the comfortable deceleration is 3.4 m/s^2. The maximum deceleration rate of car observed by various researchers varied from 1.39 m/s^2 to 3.09 m/s^2 according to the reports published from 1991 to 2005 (Maurya and Bokare, 2012). The maximum acceleration rate of both US EPA's standard 'city' and 'highway' light-duty vehicles is 1.5 m/s^2 (Le Vine et al., 2015). Consequently, we discuss the scenarios when MADR equals 1.4, 1.8, 2.2, 2.6, 3.0, and 3.4, respectively.

- Crash Potential Index (CPI)

Crash Potential Index (CPI) refers to the probability that the DRAC of a vehicle exceeds its MADR for every 0.1 s of the observation time (Cunto and Saccomanno, 2008). As shown in Fig. 4, the crash accident is deemed to occur between *Car1* and *Car2* when *Car1* 'hits' *Car2* in both X and Y directions simultaneously, i.e. $DRAC_X^{car1\,\&\,car2} > MADR_X^{car1\,\&\,car2}$ and $DRAC_Y^{car1\,\&\,Car2} > MADR_Y^{car1\,\&\,car2}$. The probability that the crash accident occurs between *Car1* and *Car2* at the timestamp t, $P^{car1\,\&\,car2}(t)$, is also assumed to follow Bernoulli distribution, which is:

$$P^{car1\,\&\,car2}(t) = \begin{cases} 1, & if\ P_X^{car1\,\&\,car2}(t) = P_Y^{car1\,\&\,car2}(t) = 1 \\ 0, & Otherwise \end{cases} \tag{11}$$

The general probability that the crash accident occurs between two cars, which refers to CPI, for the i^{th} LC event can be obtained as:

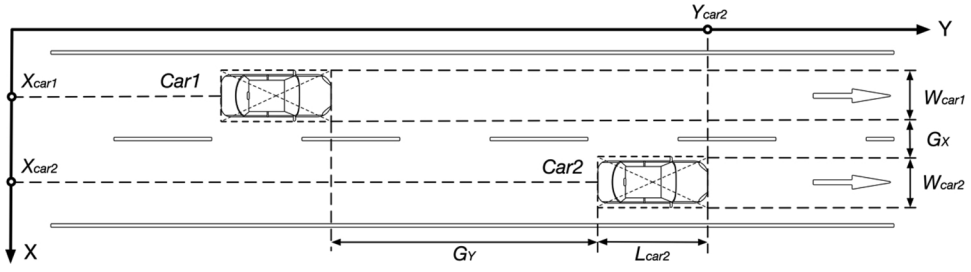


Fig. 4. Decomposition of gap between cars.

Table 1
Candidate features.

Information/behavior groups		Explanation	Statistical descriptions	Number of features
Car information	Car length	The lengths of the five cars in an LC	(constant)	5
	Car width	The widths of the five cars in an LC	(constant)	5
General behaviors	LC frequency	The LC time(s) observed on the road segment	(constant)	1
	Location	The location labeled with the X, Y coordinates of the front-center of the car	S[1]	130
	Velocity	The velocity in the traveling direction	S	65
	Acceleration	The acceleration in the traveling direction	S	65
	Space headway	The distance between the front-centers of a car and its preceding car	S	65
	Time headway	The time travel from the front-center of a car to the front-center of its preceding car at the current speed	S	65
LC behaviors	LC duration	The duration of an LC event	(constant)	1
	Lane IDs	The lane IDs of the original and target lanes	(constant)	2
	Location	The location labeled with the X, Y coordinates of the front-center of the car	S	130
	Relative location	The location labeled with the X, Y coordinates of the front-center of a car with respect to its start location labeled with the X, Y coordinates of the front-center of the car in an LC event	S	130
	Gap	The distance gaps in the X, Y directions between *sub* and one of its surrounding cars (i.e. *pre1*, *pre2*, *fol1, and fol2*)	S	104
	Velocity	The velocity in the traveling direction	S	65
	Acceleration	The acceleration in the traveling direction	S	65
	Decomposed velocity	The velocities decomposed in the X, Y directions	S	130
	Decomposed acceleration	The accelerations decomposed in the X, Y directions	S	130
	Velocity difference	The velocity differences in the X, Y directions between each pair[2] of cars	S	260
	Acceleration difference	The acceleration differences in the X, Y directions between each pair of cars	S	260
	Correlation of relative locations	The correlation between the relative locations of each pair of cars	Correlation coefficient and covariance	40
	Correlation of gaps	The correlation between each pair of the gaps between *sub* and its surrounding cars	Correlation coefficient and covariance	24
	Correlation of velocities	The correlation between the decomposed velocities of each pair of cars	Correlation coefficient and covariance	40
	Correlation of accelerations	The correlation between the decomposed accelerations of each pair of cars	Correlation coefficient and covariance	40

1.S refers to the set of statistical descriptions comprised of mean, max, min, range, 0.05 quantiles, 0.25 quantiles, 0.5 quantiles (= median), 0.75 quantiles, 0.95 quantiles, kurtosis, skewness, standard deviation, and variable coefficient.
2.There are 10 car-car pairs in an LC event, i.e., sub & pre1, sub & pre2, sub & fol1, sub & fol2, pre1 & pre2, pre1 & fol1, pre1 & fol2, pre2 & fol1, pre2 & fol2, and fol1 & fol2.

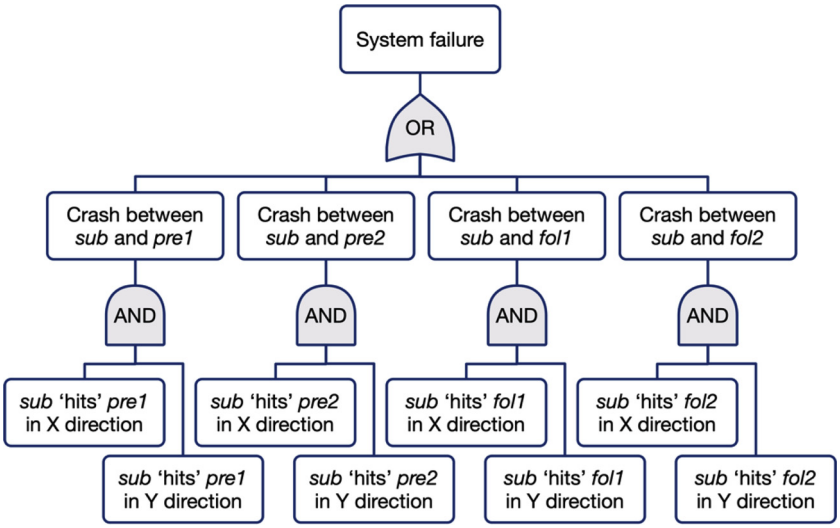


Fig. 5. Fault tree of LC.

$$P^{car1\,\&\,car2} = CPI^{car1\,\&\,car2} = \frac{\sum_{t=0}^{T} P^{car1\,\&\,car2}(t)\cdot\Delta t}{T} \tag{12}$$

where T refers to the LC duration.

3.4.2. Fault tree analysis

Fault tree analysis is a prominent method to analyze the risks related to safety or economic crisis (Ruijters and Stoelinga, 2015). In this study, we define an LC event that involves five cars shown in Fig. 2 as a system. Fig. 5 shows the structure of the fault tree designed for an LC event. The relation between the event at the middle level and the sub-event at the bottom level is described by Eqs 9, 10 and 11. The system is deemed to fail if the crash accident occurs between *sub* and any one of its surrounding cars in an LC event. The LC risk is defined as the probability of system failure. The probability of system failure (Ruijters and Stoelinga, 2015) for an LC event, P_F, can be obtained as:

$$P_F = 1 - \prod_{n}(1 - P^{sub\,\&\,pre1}) = 1 - (1 - P^{sub\,\&\,pre1})(1 - P^{sub\,\&\,pre2})(1 - P^{sub\,\&\,fol1})(1 - P^{sub\,\&\,fol2}) \tag{13}$$

3.4.3. Risk classification

Regression (for continuous outputs) and classification (for discrete outputs) are two forms of supervised learning which are important constituents of machine learning (Rasmussen, 2004). However, a disadvantage of using the regression approach is the high computational cost, especially when the number of features is larger (Meyer and Booker, 2001). Considering the large number of features in this study, we employ classification learning to select key features and predict LC risk level (the details will be discussed in Section 3.5.2). We apply k-Means algorithm, which is an unsupervised clustering method (Cover and Hart, 1967), to cluster the LC risk and classify the LC events into *n* risk levels, which are defined as the label in preparation for classification learning.

3.5. Feature selection and risk prediction

3.5.1. Resampling methods

Imbalanced dataset refers to the dataset suffering from class imbalance problem which means that one or some of the classes contain many more samples than others (Guo et al., 2016). The minority class may even not be detected when an imbalanced dataset is trained (Nekooeimehr and Lai-Yuen, 2016). Four commonly-used resampling methods, Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), Repeated Edited Nearest Neighbor (RENN) (Tomek, 1976), SMOTEENN (Batista et al., 2004), and SMOTETomek (Wilson, 1972;Tomek, 1976) are evaluated as a technique to overcome the class imbalance problem in this study. Considering the multi-class scenario in this study, we integrate one-against-all (OAA) scheme (Rifkin and Klautau, 2004) and the resampling methods to address the class imbalance problem.

3.5.2. Supervised learning method

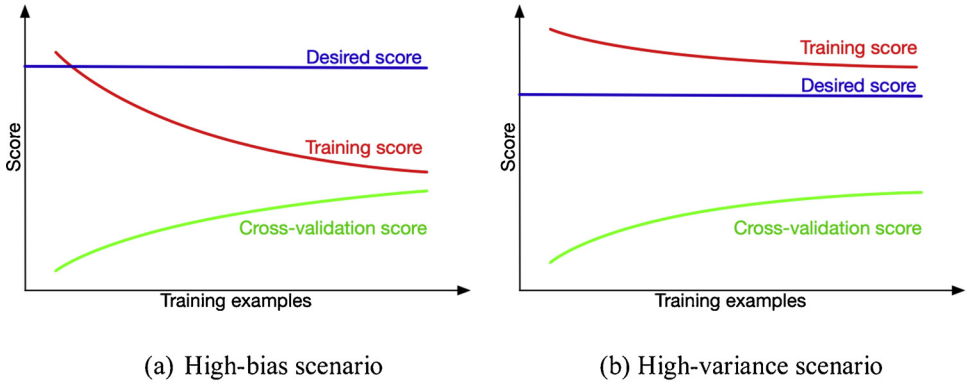
Classifier is a supervised classification learning technique that takes the values of various features of an example and predicts the class label that the example belongs to (Pereira et al., 2009). Once trained, the classifier can determine how important the information that the features contain is to the label of the example, and this relationship is tested on the test dataset. In this study, Random Forest (RF) classifier is applied as the learning model to select key features and predict LC risk. RF uses bagging, which is a bootstrap aggregation technique, to build an ensemble of decision trees as base classifiers (Breiman, 2001). RF has better performance in classification task compared to other classifiers, and it has several advantages such as excellent predictive performance, high computational efficiency, etc. (Díaz-Uriarte and De Andres, 2006; Chutia et al., 2016).

3.5.3. Model evaluation

The evaluation of the learning model trained with the resampled dataset is proposed from two perspectives, fitness performance (i.e., to judge whether the model is overfitting or underfitting) and prediction performance. Underfitting occurs when a learning model cannot sufficiently discriminate the noteworthy regularities from training examples. Overfitting occurs when a complex model is over-trained to memorize a particular set of training examples (Bishop, 2006). In this study, the fitness performance is evaluated by learning curves, and the prediction performance is measured by the evaluation measures such as precision, recall, and F1-score. The purpose of the evaluation is to find the most suitable resampling method to reduce the negative effect caused by the imbalanced dataset.

a Learning curves

Learning curves can represent how fast a learning model is improved with increasing size of training examples (Amari, 1993), and show the relations among learning accuracy, model complexity, and the size of training examples (Amari and Murata, 1993). Learning curves show the expectation values of the training and test accuracy as a



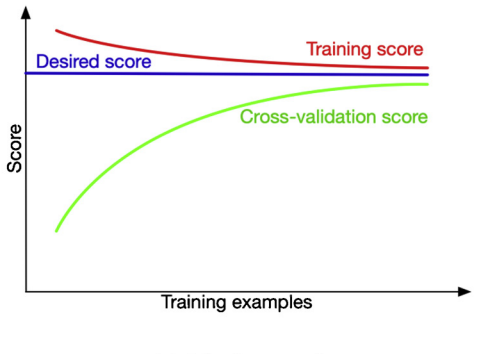


Fig. 6. Examples of learning curves.

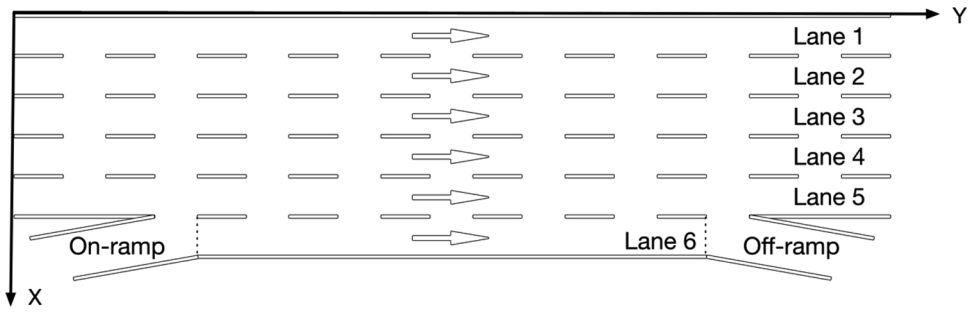


Fig. 7. Schematic illustration of study road segment.

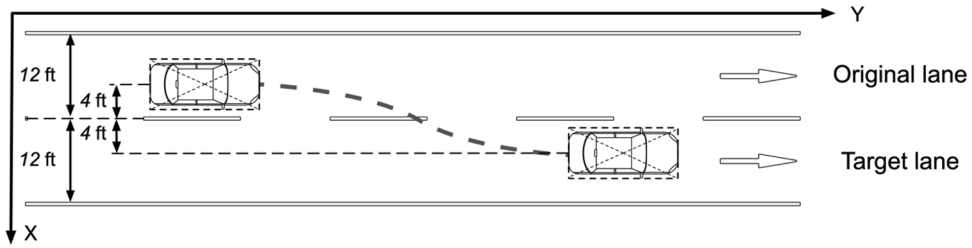


Fig. 8. Illustration of LC process.

function of training examples (Cortes et al., 1994) and provide an opportunity to diagnose the potential high-bias or high-variance in a learning process (Kohavi, 1995). High-bias in the learning process would lead to underfitting, while high-variance would result in overfitting (Domingos, 2012). Consequently, learning curves are capable of judging whether the model is underfitting or overfitting.

In this study, we apply the training score and cross-validation score (Blockeel and Struyf, 2002) to represent training accuracy and test accuracy. The following explains how to judge whether the model is underfitting or overfitting from learning curves. As shown in Fig. 6, learning curves refer to the curves of training score and cross-validation score. Fig. 6(a) indicates a high-bias scenario where both learning curves converge toward an undesired low score. Fig. 6(b) indicates a high-variance scenario where there is a large gap between the two learning curves. Fig. 6(c) indicates an ideal scenario where both learning curves converge at a high score. Consequently, the learning curves in Fig. 6(a) and Fig. 6(b) show underfitting and overfitting, respectively.

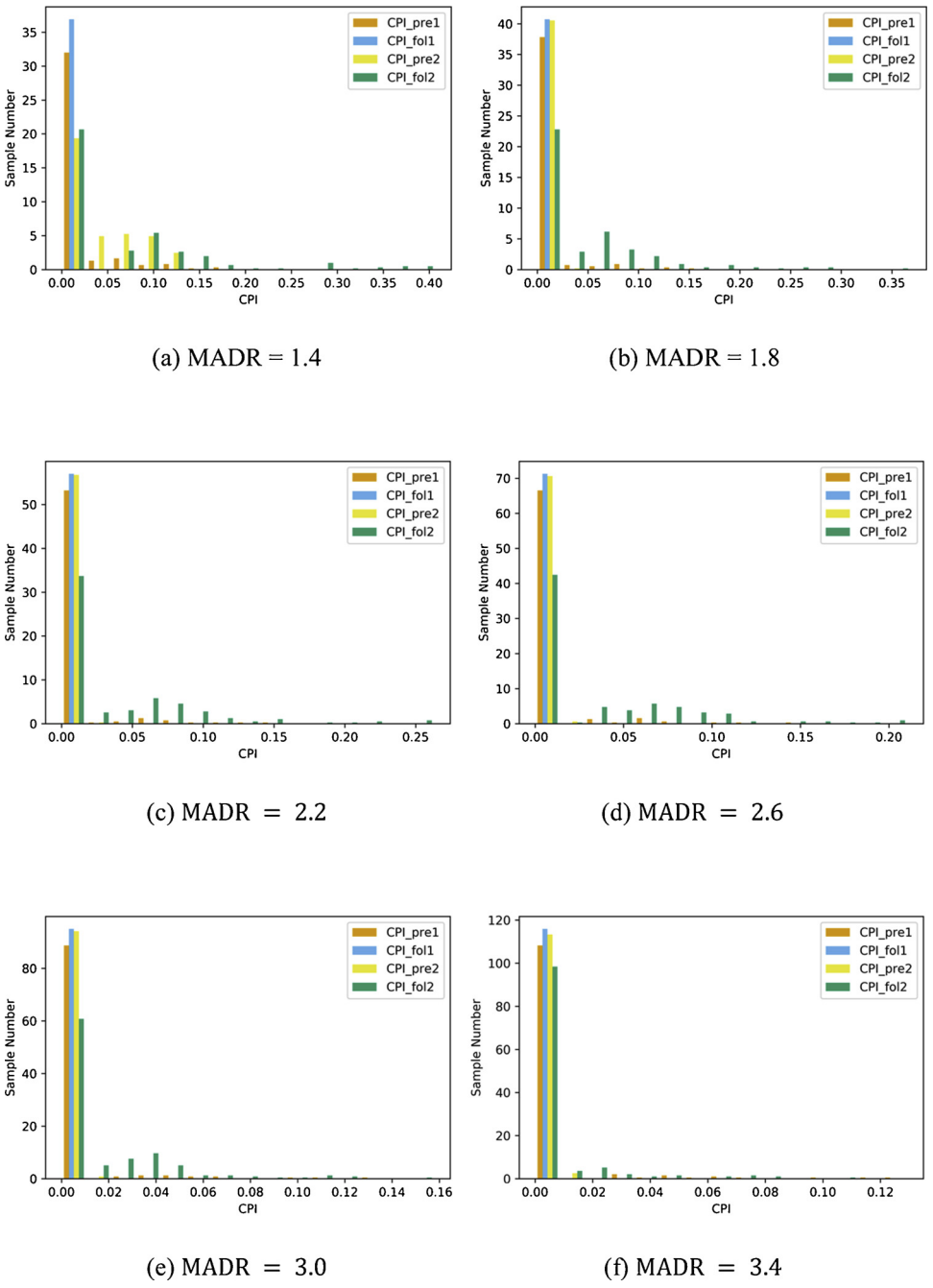


Fig. 9. Histograms of CPIs.

- Evaluation measures

In this study, we use the evaluation measures of precision (P), recall (R), and F1 score (F1) to measure the prediction (i.e. test) performance. K-fold cross-validation with k equal to 5, the recommended value to measure prediction performance (Rodriguez et al., 2010), is conducted to evaluate the learning model for each resampled dataset. The evaluation measures are applied to assess the prediction accuracy (or error) of the cross-validation. The functions of the evaluation measures are listed as follows (Powers, 2011):

$$P = \frac{TP}{TP + FP} \tag{14}$$

$$R = \frac{TP}{TP + FN} \tag{15}$$

$$F1 = \frac{2(P \cdot R)}{P + R} \tag{16}$$

where TP refers to the examples correctly predicted as positives, FP refers to the negative examples incorrectly predicted as positives, and FN refers to the positive examples incorrectly predicted as negatives.

4. Application and evaluation

4.1. Data source

The vehicle trajectory data was provided by the Federal Highway Administration's (FHWA) Next Generation Simulation (NGSIM) project (FHWA, 2005). In this study, we employ NGSIM US-101 dataset collected from 7:50 a.m. to 8:05 a.m. on June 15, 2005. The NGSIM US-101 data was extracted from the video footages collected on a road

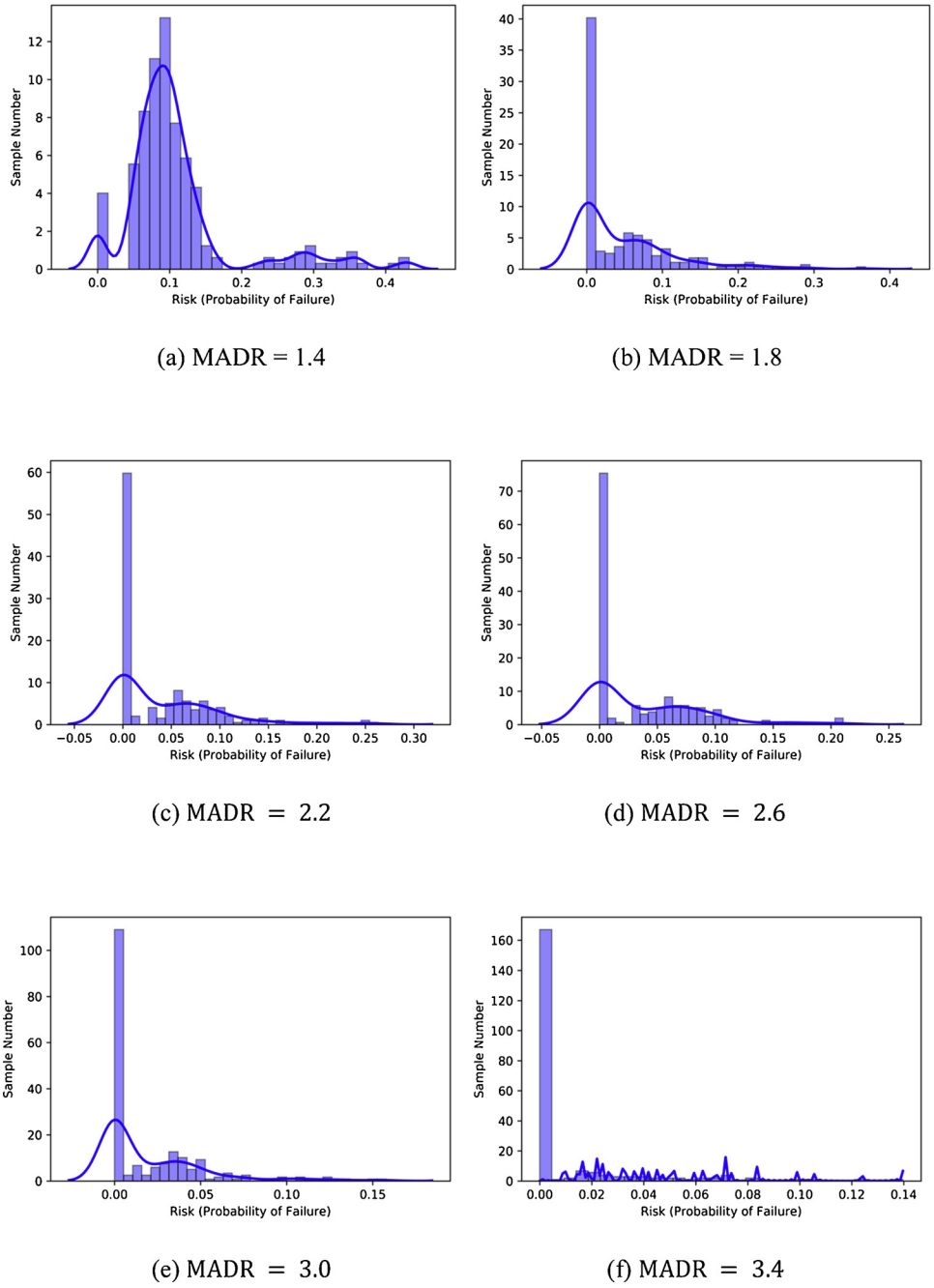


Fig. 10. Histograms of risk distribution.

Table 2
Interval partitions of LC risk classifications.

n	Level1	Level2	Level3	Level4	Level5	Level6
3	[0.000,0.082)	[0.082,0.191)	[0.191, ∞)			
4	[0.000,0.084)	[0.084,0.191)	[0.191,0.328)	[0.328, ∞)		
5	[0.000,0.024)	[0.024,0.093)	[0.093,0.093)	[0.093,0.191)	[0.328, ∞)	
6	[0.000,0.024)	[0.024,0.083)	[0.083,0.116)	[0.116,0.191)	[0.191,0.328)	[0.328, ∞)

segment of US Highway 101 in Los Angeles, California, USA. NGSIM project collected the trajectory data of cars, motorcycles, and trucks on the road segment. The dataset provides the motion information (e.g. location, velocity, acceleration, etc.) of each vehicle on the road segment every 0.1 s ($\Delta t = 0.1s$).

The road segment consists five mainline lanes and one auxiliary lane which is a portion of the corridor that connects the on-ramp and off-ramp. In this study, we only consider the LC events that occur on the six abovementioned lanes except for on-ramp and off-ramp. As shown in Fig. 7, in this study, lateral (X) coordinate is defined with respect to the left-most edge of the road segment in the direction of travel, and longitudinal (Y) coordinate is defined with respect to the entry edge of

Table 3
Resampled sample sizes of LC risk levels.

n	Resampling methods	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
3	No method	85	116	24			
	SMOTE	116	116	116			
	RENN	15	26	24			
	SMOTEENN	50	16	83			
	SMOTETomek	107	106	111			
4	No method	89	112	15	9		
	SMOTE	112	112	112	112		
	RENN	22	16	9	15		
	SMOTEENN	36	12	83	71		
	SMOTETomek	105	100	109	106		
5	No method	13	97	91	15	9	
	SMOTE	97	97	97	97	97	
	RENN	9	2	15	13	13	
	SMOTEENN	93	11	8	58	58	
	SMOTETomek	97	79	87	92	90	
6	No method	13	72	81	35	15	9
	SMOTE	81	81	81	81	81	81
	RENN	13	72	2	35	15	9
	SMOTEENN	74	11	1	27	52	43
	SMOTETomek	80	74	73	74	79	76

the road segment in the direction of travel. Lane 1 is the farthest left lane, and the lane 6 (the auxiliary lane) is the farthest right lane.

The lane width of the freeway in the US is 12 ft (3.6 m) (FHWA, 2014). As shown in Fig. 8, we define that an LC event starts and ends when the distance between the front-center of the LC car and the dotted lane marking between the original lane and target lane is 4 ft. Accordingly, we extract 225 LC events on the road segment for study, with each event involving a cluster of 5 cars.

4.2. Sensitivity analysis

In this section, we conduct the sensitivity analysis of CPIs and LC risk with different MADR values. Each LC event in this study is regarded as a sample. Fig. 9 presents the distribution of CPI when MADR equals 1.4, 1.8, 2.2, 2.6, 3.0, and 3.4. CPI_pre1, CPI_fol1, CPI_pre2, and CPI_fol2 refer to the CPIs (i.e. the probability of the crash accident) of *sub* and its surrounding cars. From Fig. 9, we find that CPI_fol2 and CPI_fol1 are most likely to be the highest one and the lowest one among the four CPIs in an LC event. This means that the crash accident potential between the LC car and its following car in the target lane is higher than that between the LC car and the other three surrounding cars, while the potential between the LC car and its following car in the original lane is lower than the others. Additionally, we found that the number of the samples that fall in the lowest CPI interval decreases gradually with the decrease of MADR. The finding indicates that the risk quantification model with MADR equal to 1.4 achieves better discrimination of the four CPIs than the models with MADR equal to the other values.

Figs. 10 illustrates the distribution of LC risk (i.e., the probability of system failure) when MADR equals 1.4, 1.8, 2.2, 2.6, 3.0, and 3.4. The blue curve refers to the density curve in each histogram. We found that the risk values of most samples are equal to zero when MADR equals 1.8, 2.2, 2.6, 3.0, and 3.4, which means that the majority of the LC events extracted from NGSIM dataset are safe. However, it is difficult to classify the samples into several risk levels and distinguish the risk of different samples when the risk values of the majority of the samples are equal to zero. Additionally, when MADR equals 1.4, the majority of the samples fall within the risk interval between 0.0 and 0.2, which achieves better discrimination of the risk distribution. Consequently, according to the sensitivity analysis of CPI and LC risk, we select MADR value of 1.4 for this study.

4.3. Resampling method selection

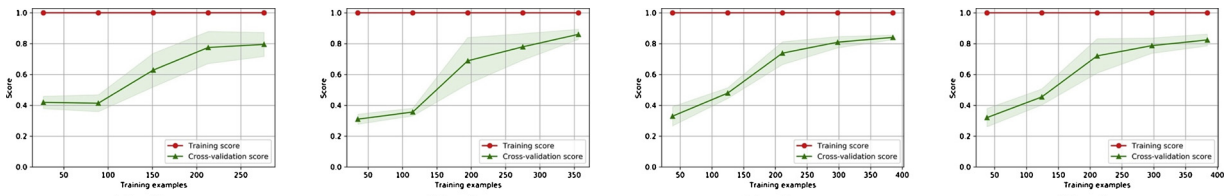
4.3.1. Labeling and resampling

We only take the classifications with n equal to 3, 4, 5, and 6 into consideration. On one hand, the classification becomes coarser and less convincing as n becomes smaller. On the other hand, the samples at certain risk level are too few to be resampled through the resampling methods if n is larger than 6. Additionally, the prediction accuracy of the classifier (i.e. RF in this study) becomes lower as n is increased to a large number (e.g. n>6). Meanwhile, the classifier is more likely to be underfitting especially when the size of the training dataset is small. The LC risk classifications resulting from k-Means with n equal to 3, 4, 5, and 6 are listed in Table 2 and the resampling results are listed in Table 3.

4.3.2. Fitness performance

As shown in Figs. 11–14, the training scores are always equal to 1.0 and the learning curves always show no underfitting regardless of the resampling methods, which is due to the RF having the capability to fit with the training samples perfectly (Izmirlian, 2004). Additionally, the training examples in the dataset can be used more than once to build decision trees, which reduces the correlations between the trees and make the RF more robust to the variations in input data and less sensitive to overfitting (Mellor et al., 2015). Consequently, severe overfitting (if it occurs) is mainly caused by the resampling method in this study. Accordingly, we can diagnose whether the resampling method can lead to overfitting from the learning curves.

From Figs. 11–14, we find that increasing the training examples can generally achieve better performance for the four models. For the combination of SMOTEENN and RF, when the training example size equals 80% of the resampled population, the gap between the learning curves narrows with the increase of n, which indicates that the model achieves the best performance when n is equal to 6. For the combination of SMOTE and RF and the combination of SMOTETomek and RF, when training example size equals 80% of the resampled population, the gaps between the learning curve become the shortest when n is equal to 4, and the models achieve the best performance. The learning curves of all the combinations of resampling method and RF converge to a high score except for the combination of RENN and RF. The gap between the learning curves of the combination of RENN and RF becomes large when n is larger than 3. Especially when n is equal to 6, the gap between the learning curves becomes the largest among the four cases, which indicates severe overfitting.



(a) n=3 (b) n=4 (c) n=5 (d) n=6

Fig. 11. Learning curves (SMOTE + RF).

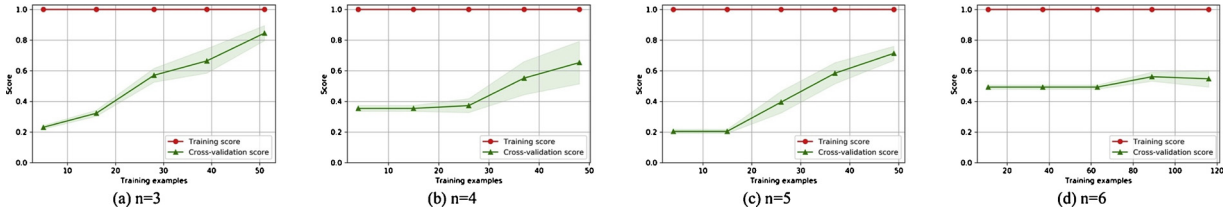


Fig. 12. Learning curves (RENN + RF).

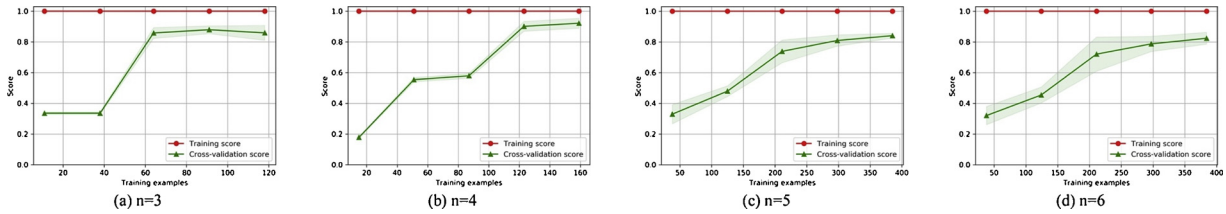


Fig. 13. Learning curves (SMOTEENN + RF).

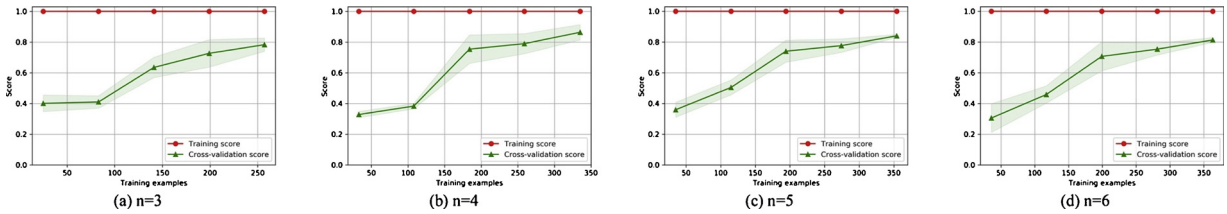


Fig. 14. Learning curves (SMOTETomek + RF).

4.3.3. Prediction performance

In this section, micro-average, macro-average (Yang, 1999), and weighted average (McCowan et al., 2004) are applied to measure the general prediction performance over the multiple risk levels. The evaluation results are shown in Table 4. Although the learning curves of the combination of SMOTEENN and RF imply that the model is less likely to be overfitting (when training example size is large), from Table 4 we find that the prediction performance is not satisfactory at certain levels of the model with different n (i.e. the performance at level 2 when n is equal to 3, the performance at level 2 when n is equal to 4, the performance at levels 2 and 3 when n is equal to 5, and the performance at level 3 when n is equal to 6). This is because the resampling results of SMOTEENN still suffer from severe class imbalance problem and the resampled sizes at certain levels are still quite small as shown in Table 4. The combination of RENN and RF also has a similar problem, which leads to poor prediction performance when n is larger than 3 as shown in Table 4. The prediction performance of both the combination of SMOTE and RF and the combination of SMOTEENN and RF reaches the highest value when n is equal to 4. Compared with the results of the combination of SMOTE and RF, the general prediction performance and the prediction performance at each level of the combination of SMOTEENN and RF are higher.

4.3.4. Summary

Although the combination of RENN and RF shows severe overfitting and undesired prediction performance when n is larger than 3, the fitness performance and the prediction performance are still satisfactory when n is equal to 3. However, as mentioned above, the classification with n equal to 3 is coarser and less convincing than the classifications with n larger than 3. Plus, the small resampled size is also highly likely to result in issues relating to generalization (Baum and Haussler, 1989). Consequently, we exclude the combination of RENN and RF. The combination of SMOTEENN and RF is also excluded due to the poor prediction performance at individual levels in spite of its desired fitness performance and satisfactory general prediction performance. Both the combination of SMOTE and RF and the combination of SMOTETomek and RF achieve their best fitness performance and prediction performance when n is equal to 4. Since the prediction performance of the combination of SMOTETomek and RF is higher than that of the combination of SMOTE and RF when n is equal to 4, we select SMOTETomek with n equal to 4 as the resampling method in this study.

4.4. Key feature selection and analysis

We classify the samples into 4 risk levels and apply SMOTETomek to resample the dataset. Then, we randomly select 70% of the samples in the resampled dataset as training examples to train RF and the other 30% samples for validation. The prediction (i.e. test) accuracy results in 0.803. We select the top 200 important features as the key features and categorize them as shown in Fig. 15. The orange plots and bins in Fig. 15 represent the features of the interactions between two cars. For example, the group *sub & pre1* contains the features of the interactions between *sub* and *pre1*, such as velocity difference, acceleration difference, etc. The red plots and bins denote the features of car's individual behaviors. For example, the group *sub* contains the features of *sub*'s behaviors, such as velocity, acceleration, etc.

From Fig. 15, we find the features of *sub* play the most important roles in LC risk determination. For the individual cars excluding *sub*, the features of *pre1* and *fol1* are more important than those of *pre2* and *fol2*, which suggests that we should pay greater attention to the individual behaviors of the surrounding cars in the original lane when analyzing LC. For the interactions between cars, the features of the interactions between *sub* and its surrounding cars are more important than those between surrounding cars. As for the interactions between surrounding cars, the group *pre2 & fol2* shows greater importance than the others, and the features which represent the correlation between the longitudinal accelerations of *pre2* and *fol2* rank higher in terms of importance. This finding implies that the interactions between the two surrounding cars in the target lane shall be of greater concern,

Table 4
Measuring results of prediction performance.

n		3			4			5			6		
Measures		P*	R*	F1*	P	R	F1	P	R	F1	P	R	F1
SMOTE + RF	Level 1	0.69	0.71	0.70	0.68	0.69	0.69	0.94	1.00	0.97	0.92	1.00	0.96
	Level 2	0.72	0.67	0.69	0.71	0.67	0.69	0.65	0.61	0.63	0.68	0.70	0.69
	Level 3	0.97	1.00	0.98	0.96	1.00	0.95	0.65	0.60	0.62	0.66	0.51	0.57
	Level 4				1.00	1.00	1.00	0.92	1.00	0.96	0.79	0.79	0.79
	Level 5							1.00	1.00	1.00	0.91	0.99	0.95
	Level 6										0.98	1.00	0.99
	Micro avg.	0.79	0.79	0.79	0.84	0.84	0.84	0.84	0.84	0.84	0.83	0.83	0.83
	Macro avg.	0.79	0.79	0.79	0.84	0.84	0.84	0.83	0.84	0.84	0.82	0.83	0.82
	Weighted avg.	0.79	0.79	0.79	0.84	0.84	0.84	0.83	0.84	0.84	0.82	0.83	0.82
RENN + RF	Level 1	0.93	0.87	0.90	0.69	0.91	0.78	0.71	0.92	0.80	0.57	0.62	0.59
	Level 2	0.85	0.88	0.87	0.61	0.69	0.65	0.67	0.92	0.77	0.56	0.92	0.69
	Level 3	0.83	0.83	0.83	0.64	0.60	0.62	0.75	0.46	0.57	0.00	0.00	0.00
	Level 4				1.00	0.11	0.20	0.68	0.87	0.76	0.29	0.11	0.16
	Level 5							1.00	0.11	0.20	0.00	0.00	0.00
	Level 6										0.00	0.00	0.00
	Micro avg.	0.86	0.86	0.86	0.66	0.66	0.66	0.70	0.70	0.70	0.53	0.53	0.53
	Macro avg.	0.87	0.86	0.87	0.74	0.58	0.56	0.76	0.66	0.62	0.24	0.27	0.24
	Weighted avg.	0.86	0.86	0.86	0.70	0.66	0.62	0.74	0.70	0.65	0.40	0.53	0.43
SMOTEENN + RF	Level 1	0.77	0.86	0.81	0.75	0.92	0.83	0.99	1.00	0.99	1.00	1.00	1.00
	Level 2	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.36	0.42	1.00	0.45	0.62
	Level 3	0.91	1.00	0.95	0.97	1.00	0.98	0.50	0.12	0.20	0.00	0.00	0.00
	Level 4				1.00	1.00	1.00	0.88	1.00	0.94	0.85	1.00	0.91
	Level 5							0.97	0.97	0.97	0.96	0.98	0.97
	Level 6										0.98	1.00	0.99
	Micro avg.	0.85	0.85	0.85	0.93	0.93	0.93	0.93	0.93	0.93	0.96	0.96	0.96
	Macro avg.	0.56	0.62	0.59	0.68	0.70	0.70	0.77	0.69	0.70	0.80	0.74	0.75
	Weighted avg.	0.77	0.85	0.80	0.88	0.93	0.90	0.91	0.93	0.92	0.96	0.96	0.96
SMOTETomek + RF	Level 1	0.67	0.68	0.68	0.75	0.72	0.74	0.93	1.00	0.97	0.90	1.00	0.95
	Level 2	0.69	0.64	0.66	0.73	0.72	0.72	0.65	0.54	0.59	0.59	0.64	0.61
	Level 3	0.95	0.99	0.97	0.96	1.00	0.72	0.65	0.59	0.61	0.58	0.42	0.49
	Level 4				0.99	1.00	1.00	0.90	1.00	0.95	0.82	0.73	0.77
	Level 5							0.96	1.00	0.98	0.89	0.99	0.93
	Level 6										0.94	1.00	0.97
	Micro avg.	0.77	0.77	0.77	0.86	0.86	0.86	0.84	0.84	0.84	0.80	0.80	0.80
	Macro avg.	0.77	0.77	0.77	0.86	0.86	0.86	0.82	0.83	0.82	0.79	0.80	0.79
	Weighted avg.	0.77	0.77	0.77	0.86	0.86	0.86	0.83	0.84	0.83	0.79	0.80	0.79

* P, R, and F1 refer to precision, recall, and F1 score, respectively.

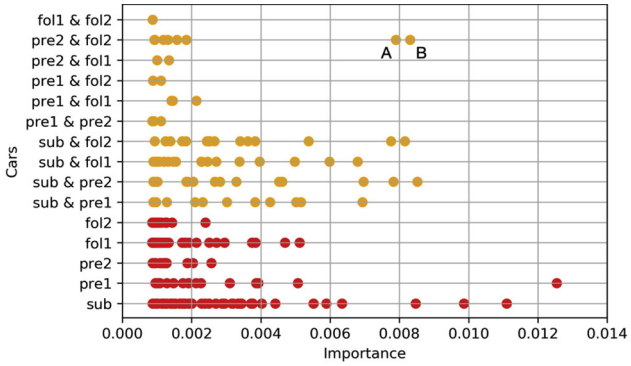
especially the acceleration features of those two cars. In comparison with the features of individual surrounding cars, we also find that features of the interactions between the *sub* and its surrounding cars show greater importance.

LC maneuver is a complex control problem for drivers, which involves situation monitoring and high-level decision making. Drivers' behaviors can be associated with vehicles' behaviors to explain the above findings. For example, drivers would shift their visual attention to the target lane almost immediately after LC onset, and draw less attention to the original lane during LC (Salvucci and Liu, 2002). The lack of attention to the original lane might result in the features of *pre1* and *fol1* being more important than those of *pre2* and *fol2* in LC risk prediction. An experiment can be designed to verify the relation between drivers' visual attention and the features' performance in the future.

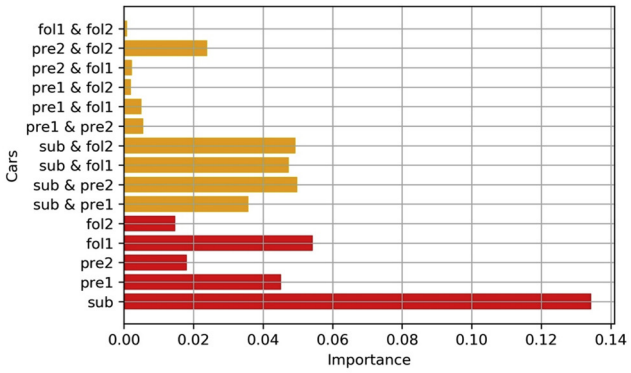
5. Conclusions

In this study, we propose a procedure for key feature selection and risk prediction for car's LC behavior. In this procedure, we define the X, Y coordinates over the selected road segment to describe car's motion behaviors. The statistical descriptions of the behaviors are defined as candidate features. CPI and fault tree are applied to quantify LC risk, and k-Means is used to cluster the LC risk and classify samples into n risk levels. Four resampling methods, SMOTE, RENN, SMOTEENN, and SMOTETomek, are proposed to solve the class imbalance problem of the LC risk dataset, and the resampled dataset is used to train and test the RF model to select key features from the candidate features and predict LC risk level. The LC events collected from NGSIM dataset are used for the application of the procedure. We evaluate the four resampling methods from the perspectives of fitness performance and prediction performance and select SMOTETomek to resample the LC risk dataset with four risk levels ($n = 4$). As a result, the mean accuracy of the risk level prediction reaches 0.803.

From the sensitivity analysis of LC risk with different MADR, we find that following cars in the original lane and target lane are respectively the safest and riskiest cars of the surrounding cars in an LC event, and a lower MADR can achieve better discrimination of risk distribution, which is in favor of the risk classification. From the evaluation of resampling methods, we find that SMOTETomek is the most suitable resampling method for the dataset in this study, which is relatively less likely to be overfitting and has higher prediction performance than the other methods. Additionally, RENN is more likely to be overfitting with increasing number of risk levels, especially for dataset with a small sample size. SMOTEENN has satisfactory general prediction performance but poor individual class performance. From the analysis of key features, we find that: a) the features of the LC car play the most important roles in LC risk determination; b) the individual behaviors of the surrounding cars in the original lane should be accorded greater attention than those of the surrounding cars in the target lane; c) the interactions between two surrounding cars in the target lane shall be of concern, especially the acceleration features of the two cars; and d) the features of the interactions between the LC car and its surrounding cars show greater importance than the features of individual surrounding cars. Those findings are significant and value-add for future research on feature analysis and risk prediction of LC.



(a) Scatter plots of feature importance (Plots A and B refer to the correlation coefficient and the covariance between the accelerations of *pre2* and *fol2* in the Y direction)



(b) Histogram of total importance

Fig. 15. Importance analysis of key features.

6. Limitations and future study

In this study, the validation of the research framework is conducted from an experimental perspective. We consider the LC scenario that involve five cars on the highway based on the vehicles' trajectory dataset collected in the US. To get a stronger validation and evaluation, more research should be conducted to gain further insight into the LC risk. For example, future study can focus on LC risk that involves more vehicle types (e.g. truck, motorcycle, etc.), LC scenarios with more than five vehicles, LC risk on different road types (e.g. urban road), and comparison of LC risk across different countries. Besides vehicle's behaviors, it shall be useful to investigate human behaviors pertaining to LC risk. The research framework proposed in this paper provides the building blocks in these future studies. From a technical perspective, we could explore more advanced methods to select key features and achieve higher LC risk prediction accuracy. Additionally, machine learning algorithm is known as a 'black-box' which can be viewed in terms of its input and output without any knowledge of its internal logic (Michie et al., 1994). Opening the 'black-box' to a 'glass-box' is of significance to the interpretability of machine learning. Consequently, in future study, we can delve deeper into key feature selection and analysis to derive a more detailed understanding of how those features affect LC risk. It is also important to improve Advanced Driver-Assistance System (ADAS) calibrated on applicable key features to remediate the occurrence of risky LC events.

Acknowledgements

This paper presents a part of the first author's PhD research.

References

AASHTO, 2001. Policy on Geometric Design of Highways and Streets. American Association of State Highway and Transportation Officials, Washington, DC.
Amari, S.I., 1993. A universal theorem on learning curves. Neural Netw. 6, 161–166.
Amari, S.I., Murata, N., 1993. Statistical theory of learning curves under entropic loss criterion. Neural Comput. 5, 140–153.
Arbis, D., Dixit, V.V., 2019. Game theoretic model for lane changing: incorporating conflict risks. Accid. Anal. Prev. 125, 158–164.
Balal, E., Cheu, R.L., Sarkodie-gyan, T., 2016. A binary decision model for discretionary lane changing move based on fuzzy inference system. Transp. Res. Part C Emerg. Technol. 67, 47–61.
Batista, G.E., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. Acm Sigkdd Explor. Newsl. 6, 20–29.
Baum, E.B., Haussler, D., 1989. What size net gives valid generalization? Adv. Neural Inf. Process. Syst. 81–90.
Bishop, C.M., 2006. Information Science and Statistics. Pattern Recognition and Machine Learning.
Blockeel, H., Struyf, J., 2002. Efficient algorithms for decision tree cross-validation. J. Mach. Learn. Res. 3, 621–650.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Butakov, V.A., Ioannou, P., 2015. Personalized driver/vehicle lane change models for ADAS. IEEE Trans. Veh. Technol. 64, 4422–4431.
Chai, C., Shi, X., Wong, Y.D., 2017. Safety Evaluation of Vehicle-to-vehicle (V2V) Communications System on Motorcycle-vehicle Interaction Based on Fuzzy Cellular Automata (FCA). No. 17-01099. .
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Chutia, D., Bhattacharyya, D., Sarma, K.K., Kalita, R., Sudhakar, S., 2016. Hyperspectral remote sensing classifications: a perspective survey. Trans. Gis 20, 463–490.
Cooper, D.F., Ferguson, N., 1976. Traffic Studies at T-Junctions. 2. A Conflict Simulation Record. Traffic Engineering and Control, pp. 17.
Cortes, C., Jackel, L.D., Solla, S.A., Vapnik, V., Denker, J.S., 1994. Learning curves: asymptotic values and rate of convergence. Adv. Neural Inf. Process. Syst. 327–334.
Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13, 21–27.
Cunto, F., Saccomanno, F.F., 2008. Calibration and validation of simulated vehicle safety performance at signalized intersections. Accid. Anal. Prev. 40, 1171–1179.

Díaz-uriarte, R., De Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3.
Domingos, P., 2012. A few useful things to know about machine learning. Commun. ACM 55, 78–87.
FHWA, 2005. Next Generation Simulation (NGSIM). Accessed September 2018. Available:. https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm.
FHWA, 2014. Safety. Accessed September 2018. https://safety.fhwa.dot.gov/geometric/pubs/mitigationstrategies/chapter3/3_lanewidth.cfm.
Gal, A., Mandelbaum, A., Schnitzler, F., Senderovich, A., Weidlich, M., 2017. Traveling time prediction in scheduled transportation with journey segments. Inf. Syst. 64, 266–280.
Guo, F., Wotring, B.M., Antin, J.F., 2010. Evaluation of Lane Change Collision Avoidance Systems Using the National Advanced Driving Simulator (No. HS-811 332).
Guo, H., Li, Y., JEnnifer, S., Gu, M., Huang, Y., Gong, B., 2017. Learning from class-imbalanced data: review of methods and applications. Expert Syst. Appl. 73, 220–239.
Guo, H., Li, Y., Li, Y., Liu, X., Li, J., 2016. BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification. Eng. Appl. Artif. Intell. 49, 176–193.
Guo, L., Ge, P.S., Yue, M., Zhao, Y.B., 2014. Lane changing trajectory planning and tracking controller design for intelligent vehicle running on curved road. Math. Probl. Eng. 2014.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
Hou, Y., Edara, P., Sun, C., 2014. Modeling mandatory lane changing using Bayes classifier and decision trees. IEEE Trans. Intell. Transp. Syst. 15, 647–655.
Hou, Y., Edara, P., Sun, C., 2015. Situation assessment and decision making for lane change assistance using ensemble learning methods. Expert Syst. Appl. 42, 3875–3882.
Izmirlian, G., 2004. Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. Ann. N. Y. Acad. Sci. 1020, 154–174.
Izquierdo, R., Parra, I., Muñoz-bulnes, J., Fernández-llorca, D., Sotelo, M., 2017. Vehicle trajectory and lane change prediction using ANN and SVM classifiers. 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC) 1–6.
Jahangiri, A., Rakha, H.A., 2015. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. IEEE trans. Intell. Transp. Syst. 16, 2406–2417.
Keyvan-ekbatani, M., Knoop, V.L., Daamen, W., 2016. Categorization of the lane change decision process on freeways. Transp. Res. Part C Emerg. Technol. 69, 515–526.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai 1137–1145.
Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Prog. Artif. Intell. 5, 221–232.
Le Vine, S., Zolfaghari, A., Polak, J., 2015. Autonomous cars: the tension between occupant experience and intersection capacity. Transp. Res. Part C Emerg. Technol. 52, 1–14.
Li, K., Wang, X., Xu, Y., Wang, J., 2016. Lane changing intention recognition based on speech recognition models. Transp. Res. Part C Emerg. Technol. 69, 497–514.
Luo, Y., Xiang, Y., Cao, K., Li, K., 2016. A dynamic automated lane change maneuver based on vehicle-to-vehicle communication. Transp. Res. Part C Emerg. Technol. 62, 87–102.
Ma, X., Yu, H., Wang, Y., Wang, Y., 2015. Large-scale transportation network congestion evolution prediction using deep learning theory. PLoS One 10, e0119044.
Mahmud, S.S., Ferreira, L., Hoque, M.S., Tavassoli, A., 2017. Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs. IATSS Res. 41, 153–163.
Maurya, A.K., Bokare, P.S., 2012. Study of deceleration behaviour of different vehicle types. Int. J. Traffic Transp. Eng. 2.
McCowan, I.A., Moore, D., Dines, J., Gatica-perez, D., Flynn, M., Wellner, P., Bourlard, H., 2004. On the Use of Information Retrieval Measures for Speech Recognition Evaluation. Idiap-RR Idiap-RR-73-2004, IDIAP, Martigny, Switzerland, pp. 0.
Mellor, A., Boukir, S., Haywood, A., Jones, S., Sensing, R., 2015. Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. ISPRS J. Photogramm. 105, 155–168.
Nekooeimehr, I., Lai-yuen, S.K., 2016. Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets. Expert Syst. Appl. 46, 405–416.
Meyer, M.A., Booker, J.M., 2001. Eliciting and Analyzing Expert Judgment, A Practical Guide. Siam.
Michie, D., Spiegelhalter, D.J., Taylor, C.C., 1994. Machine learning. Neural and Statistical Classification. 13. .
Nilsson, J., Silvlin, J., Brannstrom, M., Coelingh, E., Fredriksson, J., 2016. If, when, and how to perlane change maneuvers on highways. IEEE Intell. Transp. Syst. Mag. 8, 68–78.
Oh, C., Kim, T., 2010. Estimation of rear-end crash potential using vehicle trajectory data. Accid. Anal. Prev. 42, 1888–1893.
Park, H., Oh, C., Moon, J., Kim, S., 2018. Development of a lane change risk index using vehicle trajectory data. Accid. Anal. Prev. 110, 1–8.
Peng, Y., Abdel-aty, M., Shi, Q., Yu, R., 2017. Assessing the impact of reduced visibility on traffic crash risk using microscopic data and surrogate safety measures. Transp. Res. Part C Emerg. Technol. 74, 295–305.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, S199–S209.
Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J. Mach. Learn. Technol. 2, 37–63.
Rahman, M.S., Abdel-aty, M., 2018. Longitudinal safety evaluation of connected vehicles' platooning on expressways. Accid. Anal. Prev. 117, 381–391.
Rasmussen, C.E., 2004. Gaussian processes in machine learning. Advanced Lectures on Machine Learning. Springer, Berlin, Heidelberg.
Rifkin, R., Klautau, A., 2004. In defense of one-vs-all classification. J. Mach. Learn. Res. 5, 101–141.
Rodriguez, J.D., Perez, A., Lozano, J.A., 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32, 569–575.
Ruijters, E., Stoelinga, M., 2015. Fault tree analysis: a survey of the state-of-the-art in modeling, analysis and tools. Comput. Sci. Rev. 15, 29–62.
Salvucci, D.D., Liu, A., 2002. The time course of a lane change: driver control and eye-movement behavior. Transp. Res. Part F: Traffic Psychol. Behav. 5 (2), 123–132.
Schlechtriemen, J., Wirthmueller, F., Wedel, A., Breuel, G., Kuhnert, K.D., 2015. When will it change the lane? A probabilistic regression approach for rarely occurring events. Intelligent Vehicles Symposium (IV) 1373–1379.
Shi, X., Wong, Y.D., Li, M.Z.F., Chai, C., 2018a. Key risk indicators for accident assessment conditioned on pre-crash vehicle trajectory. Accid. Anal. Prev. 117, 346–356.
Shi, X., Wong, Y.D., Li, M.Z.F., Chai, C., 2018b. Accident Risk Prediction Based on Driving Behavior Feature Learning Using Cart and XGBoost. No. 18-06270. .
Shi, X., Wong, Y.D., Li, M.Z.F., Chandrasekar, P., Chai, C., 2019. A feature learning approach based on XGBoost for driving assessment and risk prediction. Accid. Anal. Prev. https://doi.org/10.1016/j.aap.2019.05.005.
Tomek, I., 1976. Two modifications of CNN. Trans. Syst. Man Cybern. 6, 769–772.
Wang, C., Sun, Q., Fu, R., Li, Z., Zhang, Q., 2018. Lane change warning threshold based on driver perception characteristics. Accid. Anal. Prev. 117, 164–174.
Wang, M., Hoogendoorn, S.P., Daamen, W., Van Arem, B., Happee, R., 2015. Game theoretic approach for predictive lane-changing and car-following control. Transp. Res. Part C Emerg. Technol. 58, 73–92.
Wilson, D., 1972. Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. Syst. Man Cybern. 408–421.
Woo, H., Ji, Y., Kono, H., Tamura, Y., Kuroda, Y., Sugano, T., Yamamoto, Y., Yamashita, A., Asama, H., 2017. Lane-change detection based on vehicle-trajectory prediction. IEEE Robot. Autom. Lett. 2, 1109–1116.
Yan, F., Eilers, M., Baumann, M., Luedtke, A., 2016. Development of a lane change assistance system adapting to driver's uncertainty during decision-making. Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications 93–98.
Yang, D., Zheng, S., Wen, C., Jin, P.J., Ran, B., 2018. A dynamic lane-changing trajectory planning model for automated vehicles. Transp. Res. Part C Emerg. Technol. 95, 228–247.
Yang, Y., 1999. An evaluation of statistical approaches to text categorization. Inf. Retr. 1, 69–90.
Zhao, P., Lee, C., 2018. Assessing rear-end collision risk of cars and heavy vehicles on freeways using a surrogate safety measure. Accid. Anal. Prev. 113, 149–158.
Zheng, J., Suzuki, K., Fujita, M., 2014. Predicting driver's lane-changing decisions using a neural network model. Simul. Model. Pract. Theory 42, 73–83.
Zheng, Z., 2014. Recent developments and research needs in modeling lane changing. Transp. Res. Part B Methodol. 60, 16–32.
Zheng, Z., Ahn, S., Chen, D., Laval, J., 2013. The effects of lane-changing on the immediate follower: anticipation, relaxation, and change in driver characteristics. Transp. Res. Part C Emerg. Technol. 26, 367–379.
Zhou, B., Wang, Y., Yu, G., Wu, X., 2017. A lane-change trajectory model from drivers' vision view. Transp. Res. Part C Emerg. Technol. 85, 609–627.