Hindawi
Discrete Dynamics in Nature and Society
Volume 2018, Article ID 4954621, 12 pages
https://doi.org/10.1155/2018/4954621

Research Article

Vehicle Driving Risk Prediction Based on Markov Chain Model

Xiaoxia Xiong, Long Chen, and Jun Liang

School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China

Correspondence should be addressed to Long Chen; chenlong@ujs.edu.cn

Received 18 November 2017; Accepted 20 December 2017; Published 18 January 2018

Academic Editor: Xiaohua Ding

A driving risk status prediction algorithm based on Markov chain is presented. Driving risk states are classified using clustering techniques based on feature variables describing the instantaneous risk levels within time windows, where instantaneous risk levels are determined in time-to-collision and time-headway two-dimension plane. Multinomial Logistic models with recursive feature variable estimation method are developed to improve the traditional state transition probability estimation, which also takes into account the comprehensive effects of driving behavior, traffic, and road environment factors on the evolution of driving risk status. The "100-car" natural driving data from Virginia Tech is employed for the training and validation of the prediction model. The results show that, under the 5% false positive rate, the prediction algorithm could have high prediction accuracy rate for future medium-to-high driving risks and could meet the timeliness requirement of collision avoidance warning. The algorithm could contribute to timely warning or auxiliary correction to drivers in the approaching-danger state.

1. Introduction

Road traffic injuries and deaths have been a major public health issue globally. According to World Health Organization (WHO), approximately 1.25 million people die from roadway traffic accidents each year, while 20~50 million people suffer nonfatal injuries with many resulting in disabilities [1]. Vehicle driving risk prediction based on the perception of real-time movement and environment features of the vehicle could be vital for developing collision warning/intervention strategies in intelligent driver assistance systems to reduce collision risks and improve roadway safety.

Most existing collision warning methods calculate the selected warning parameters in real time and compare them with the default thresholds of different risk levels, and the most widely used warning parameters include time-to-collision (TTC), time-headway (THW), and distance [2–4]. However, it is too simplified to describe the whole driving risk evolution process from the formation of risk to the occurrence of accident with only a single warning parameter, and more complex models and algorithms are required for more intelligent driving risk prediction. Although such advanced collision warning models/algorithms have received increasing attention over the years, many published studies usually only consider the vehicle operating characteristics (such as vehicle's relative position to potential conflicts, vehicle speed, and acceleration characteristics) [5–7], while ignoring the impact of dynamic driver behavior, road, and environmental characteristics on the driving risk status, which has been researched and confirmed in many traffic accident causation studies [8–10]. Thus, a driving risk prediction method that could reflect the dynamic changes in driver behavior, road, and environment is needed.

As the future risk state of a driving vehicle has strong randomness and no aftereffect (i.e., "the state of the previous moment has no direct influence on the state of the next moment" [11]), the driving risk evolution process follows the Markov property. Markov chains have been widely used in the engineering field and have already been applied to the area of transportation, such as traffic flow and travel speed forecasting [11–13], but have not been extensively researched in driving risk prediction. To solve the problem of driving risk prediction for vehicle collision avoidance, the paper aims to explore a Markov chain driving risk state forecasting model that considers the dynamic changes of real-time driver behavior, road, and environmental characteristics. The results of the study could provide a new basis for vehicle collision warning and risk control.

Table 1: Retrieved risk-related attributes from "100-car" NDS dataset.

		Recoded data attributes
Level 1	(1) Vehicle movement parameters	Vehicle speed, acceleration, distance from the leading vehicle, rate of change in distance from the leading vehicle
Level 2	(2) Driver attributes	Driver's driving maneuver prior to (near) accident, driver attention area, the number of secondary tasks, the highest complexity level of the secondary tasks
	(3) Road attributes	Number of lanes, traffic flow density, road alignment, road longitudinal slope
	(4) Environmental attributes	Lighting, weather conditions, roadway conditions

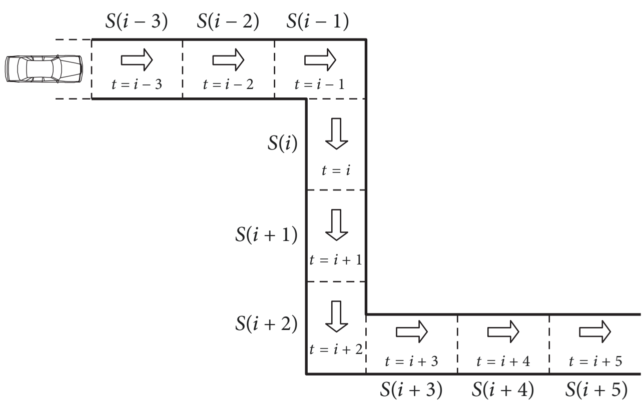


Figure 1: Vehicle driving state transition process.

The driving risk prediction problem is briefly stated in the next section, followed by a description of the driving state data for establishing the real-time risk prediction model. The methodology for model development is presented in Section 4, followed by model parameter experiment, validation, and results of data analysis. The final section summarizes the findings and concludes the paper.

2. Problem Statement

In order to realize the real-time driving risk prediction, driving state parameters of the vehicle need to be collected in real time through sensors and instruments. The time interval of data acquisition by sensors and instruments divides the continuous time variable into a discrete time series $t = \{\ldots, i-3, i-2, i-1, i, i+1, i+2, i+3, \ldots\}$. Accordingly, the continuous evolving process of vehicle driving state can be expressed as a discrete sequence corresponding to each discretized time moment: $S(t) = \{\ldots, S(i-3), S(i-2), S(i-1), S(i), S(i+1), S(i+2), S(i+3), \ldots\}$ (Figure 1) [11]. The discrete sequence $S(t)$ has strong randomness and the probability of the vehicle being at the next state depends only on the current state and not the previous states:

$$\Pr\left\{S(i+1) = S' \mid S(i), S(i-1), \ldots, S(1)\right\} = \Pr\left\{S(i+1) = S' \mid S(i)\right\}. \quad (1)$$

As such, the above-described vehicle driving state evolution process accords with the Markov property and could be well described by a Markov model.

3. Data Source

The driving risk evolution observation data in this paper were derived from the "100-car" Natural Driving Study (NDS) database collected by Virginia Tech from 2004 to 2005 [14], which includes 68 accidents and 760 near-accidents (where drivers took an emergency braking or evasive steering behavior) data. Each set of data records information concerning vehicle movement features as well as driver behavior, traffic, and environmental status from 30 s before and 10 s after the occurrence of accident (or near-accident), which could well meet the research objective of the paper in exploring driving risk prediction method that could reflect dynamic driver behavior, road, and environmental characteristics. As the dataset does not record the complete set of parameters for vehicle driving involved in lane-changing conflicts, only rear-end accidents and near-rear-end accidents samples were selected for study. After deleting the invalid observation records (with missing/unreasonable values) and recoding the selected risk-related attributes, a total of $N = 114$ samples $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$ were finally obtained, where each sample $\mathbf{X}_i$ ($i = 1, 2, \ldots, N$) is a time series recording vehicle movement status values (with a duration of T_i and sampling interval at 0.1 s or 10 Hz), along with a set of driver, road and, environment status attributes values, as summarized in Table 1.

As presented in Table 1, the four categories of the recoded attributes could be grouped into Level 1 and Level 2, where Level 1 attributes feature the moving status of the vehicle, while Level 2 attributes feature the state-changing affecting/causal factors. As such, Level 1 attributes would be first used for the classification of driving risk states in Section 4.1 and then for the estimation of state transition probability of Markov chain model together with Level 2 attributes in Section 4.2.

The obtained time-series samples were then randomly divided into two groups, of which 70 (about 60%) time-series samples would be used for model training (in Section 5) and the remaining 44 (about 40%) for model verification (in Section 6).

4. Methodology

Following the Markov chain property, the state of a research object at time $t+1$ is determined by the product of the initial state probability distribution vector at time t ($\boldsymbol{\pi}_t$) and the state transition probability matrix A [15] as follows:

$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t A. \tag{2}$$

Therefore, the paper would focus on the two aspects of research including risk state classification and transition probability estimation.

4.1. Driving Risk State Classification. Previous research has demonstrated that vehicle driving risk level could be well characterized by the driver's braking features [16–18]. Therefore, the vehicle movement parameters at the start of braking in the accident and near-accident samples were selected to classify the instantaneous driving risk level. TTC and THW are two widely recognized parameters in measuring vehicle driving risk, which are defined as follows [3]:

$$\begin{aligned} \mathrm{TTC} &= \frac{D}{V_r} \\ \mathrm{THW} &= \frac{D}{V_s}, \end{aligned} \tag{3}$$

where D measures the distance from the subject vehicle to the leading vehicle; V_r and V_s represent the relative speed of the two vehicles and the traveling speed of the subject vehicle. Equations (3) imply that the risk evaluation by TTC is limited in the small relative distance risk scenario, and THW could not account for the scenario when the relative speed of the two vehicles increases. To overcome the limits in using one of the two parameters, K-means clustering of $\{-\mathrm{iTTC}, \mathrm{THW}\}$ vectors observed at the start of braking was implemented to identify different risk level areas (K-means clustering algorithm was employed here for its wide application to and good performance in risk state classification [5, 16, 17]). Note that iTTC is the inverse of TTC and was utilized instead of TTC to avoid the problem of infinite TTC when the relative speed is very small, and the negative sign before iTTC ensures that it maintains the same increasing/decreasing trend as THW when risk level changes (i.e., both a smaller −iTTC and a smaller THW indicate a higher driving risk).

Both of the training and test sets were utilized for risk level clustering (with a total of 190 braking processes obtained), and the optimum number of clusters was estimated to be five by the elbow method [19]. The general distribution pattern of $\{-\mathrm{iTTC}, \mathrm{THW}\}$ pair clusters at the start of driver braking is presented in Figure 2 (where each cluster area is represented with a different color). As could be noted from Figure 2, the whole risk observations at braking could generally be divided into five regions following the lines near iTTC = 0.7, THW = 0.9, THW = 1.3, THW = 1.8, and THW = 2.5, respectively. Notice that there exist risk pairs with a negative iTTC (which is assigned to no risk level according to TTC-based classification) but with a relatively small THW, indicating the necessity for joint consideration of both parameters for risk assessment. According to the distribution of the clustering results, as well as referencing the typical TTC-based risk level classification according to accident-to-conflict ratio (which is estimated to be 0.8, 0.6, 0 for $1.0 \le \mathrm{iTTC}$, $0.67 \le \mathrm{iTTC} < 1.0$, $\mathrm{iTTC} < 0$ intervals, resp.) [20], an instantaneous driving risk level (RL) indexing integrating driving behavior characteristics (driver braking features) and conflict severity (accident-to-conflict ratio) could be defined as in Table 2 (also marked in Figure 2).

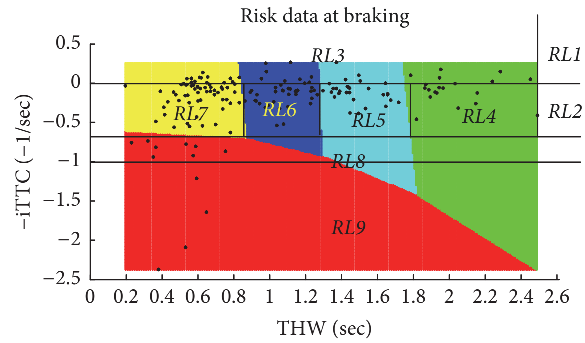


FIGURE 2: Vehicle movement parameter distribution at the start of braking.

TABLE 2: Instantaneous driving risk level definition.

Risk group	Risk level	Thresholds
High risk	RL9	$1.0 \le \mathrm{iTTC}$
	RL8	$0.67 \le \mathrm{iTTC} < 1.0$
	RL7	$0 \le \mathrm{iTTC} < 0.67$ & $\mathrm{THW} < 0.9$
⇓	RL6	$0 \le \mathrm{iTTC} < 0.67$ & $0.9 \le \mathrm{THW} < 1.3$
	RL5	$0 \le \mathrm{iTTC} < 0.67$ & $1.3 \le \mathrm{THW} < 1.8$
	RL4	$0 \le \mathrm{iTTC} < 0.67$ & $1.8 \le \mathrm{THW} < 2.5$
Low risk	RL3	$\mathrm{iTTC} < 0$ & $\mathrm{THW} < 2.5$
	RL2	$0 \le \mathrm{iTTC} < 0.67$ & $2.5 \le \mathrm{THW}$
	RL1	$\mathrm{iTTC} < 0$ & $2.5 \le \mathrm{THW}$

In order to improve model performance regarding accuracy as well as timeliness, instead of using single point risk level observation to define risk condition, driving risk state at any time t was defined over a time window (ending at time t) based on statistics to capture the range and trend of risk level over a short time period.

4.1.1. Rolling Time Window. A rolling time window scheme was used to segment the time-series data samples. As shown in Figure 3, the length of the rolling time window (TW) is w, which is continuously rolling forward at a rolling interval φ (equal to the sampling interval [0.1 s or 10 Hz] in the NDS dataset). Transition step δ defines the unit increase of the discrete moments of time $t = 1, 2, \ldots, T$, at which the Markov state transitions (e.g., the transition of risk states from time t to time $t+1$) are observed. Accordingly, a total of $N' = \sum_{i=1}^{N} (\lfloor (T_i - w)/\varphi \rfloor + 1)$ time windows could be retrieved after segmenting each time series sample $\mathbf{X}_i$ ($i = 1, 2, \ldots, N$), where T_i represents the total time length of $\mathbf{X}_i$. To achieve

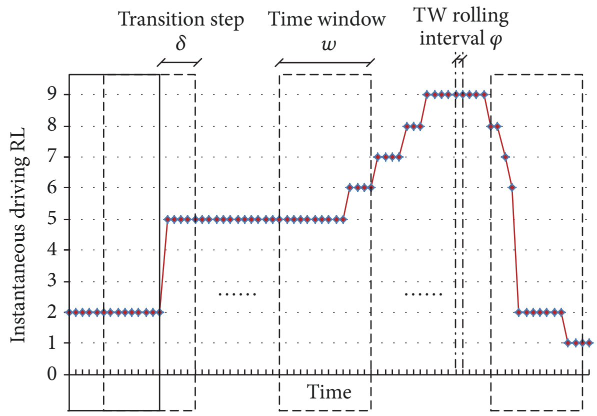


Figure 3: Diagram of rolling time window.

optimum prediction accuracy as well as timeliness, the length of rolling time window w and transition step δ for Markov chain model prediction would be calibrated based on the training sample data by grid searching (to be discussed in Section 5).

4.1.2. Time Window-Based Risk State Clustering. As described above, the driving risk state at any time t is represented by the time window ending at t and could be determined by the statistics of the observation sequence of risk levels within the corresponding time window. By principle components analysis of possible statistics, the mean risk level within time window ($\mathrm{RL}_{\mathrm{avg}}$) and a trend statistic contrast (CON) were selected as the feature statistics for each time window. The two statistics, along with the instantaneous risk level at time t ($\mathrm{RL}_{\mathrm{last}}$, i.e., the last observed risk level in time window), were then served as feature variables in risk state clustering analysis for state classification. The trend value CON is widely used in the field of image analysis to describe the variation in grayscale of an image [21]. For risk state prediction problem, CON measures the degree of variation of the risk level observation sequence within the time window, which is defined as follows:

$$\mathrm{CON} = \sum_{i,j} (j-i)\left|j-i\right| d_{ij}$$
$$d_{ij} = \frac{\text{number of risk level pairs } (i,j) \text{ with distance 1}}{\text{total number of possible pairs with distance 1}}, \quad (4)$$

where d_{ij} represents the ith row and jth column element of the risk level cooccurrence matrix (both i and j represent the instantaneous driving risk level index). Given a risk level observation sequence, the risk level cooccurrence matrix could be created by calculating how often it occurs when a contiguous risk level pair is at level i and level j. An example calculation of CON for a time window consisting of ten risk level observations is presented in Figure 4.

Formula (4) shows that CON would yield a positive value when the risk level has an increasing trend over time and a negative value when it has a decreasing trend. As such, CON not only measures the contrast intensity of each observed risk level with its neighboring observations but also reflects the characteristics of changing trend in risk level within the time window period and thus is suitable for characterizing the variation of risk level within the time window.

To facilitate future development of early collision warning scheme, driving risk state was classified into $S1$, low-risk state, $S2$, medium-risk state, and $S3$, high-risk state, by K-means clustering ($K = 3$) technique, which was performed on all the feature vectors $[\mathrm{RL}_{\mathrm{avg}}, \mathrm{RL}_{\mathrm{last}}, \mathrm{CON}]$ retrieved from the obtained time windows based on the training sample (the results of clustering would be discussed in Section 6). The classified driving risk states were defined as Markov chain states $\{S_i\}$, $i = 1, 2, 3$ for Markov chain modeling.

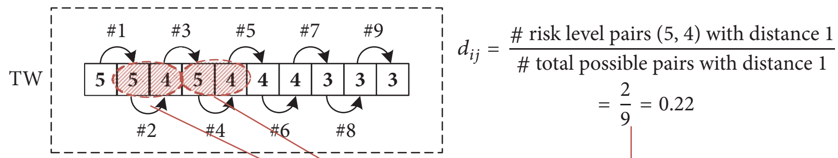
As K-means clustering is distance-based, the initial risk state probability distribution $\boldsymbol{\pi}_0$ could also be estimated based on the Euclidean distance ρ between the initially observed three-dimension feature vector $\mathbf{x}_0 = [\mathrm{RL}_{\mathrm{avg}_0}, \mathrm{RL}_{\mathrm{last}0}, \mathrm{CON}_0]$ and the risk state cluster centroids $\{\mathbf{c}_i = [\mathrm{RL}_{\mathrm{avg}_{c_i}}, \mathrm{RL}_{\mathrm{last}_{c_i}}, \mathrm{CON}_{c_i}]\}$, $i = 1, 2, 3$, as shown in the following formula:

$$\pi_{0S_i} = \frac{1/\rho(\mathbf{x}_0, \mathbf{c}_i)}{\sum_{j=1}^{K} 1/\rho\left(\mathbf{x}_0, \mathbf{c}_j\right)}, \quad i = 1, 2, 3$$
$$\rho(\mathbf{x}_0, \mathbf{c}_i) = \sqrt{\sum_{j=1}^{3} \left(\mathbf{x}_0(j) - \mathbf{c}_i(j)\right)^2}, \quad i = 1, 2, 3. \quad (5)$$

4.2. Driving Risk State Transition Probability Estimation. As shown in (2), in the Markov chain forecasting process, the estimation of state transition probability matrix would directly affect the accuracy of prediction and is the key to model development. In practical applications, the probability of state transition is usually calculated based on the transition frequency between states, which could be expressed as follows:

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1K} \\ \vdots & \ddots & \vdots \\ a_{K1} & \cdots & a_{KK} \end{bmatrix} = \begin{bmatrix} \dfrac{N(S_1, S_1)}{\sum_{k=1}^{K} N(S_1, S_k)} & \cdots & \dfrac{N(S_1, S_K)}{\sum_{k=1}^{K} N(S_1, S_k)} \\ \vdots & \ddots & \vdots \\ \dfrac{N(S_K, S_1)}{\sum_{k=1}^{K} N(S_K, S_k)} & \cdots & \dfrac{N(S_K, S_K)}{\sum_{k=1}^{K} N(S_K, S_k)} \end{bmatrix}, \quad (6)$$

where K refers to the number of classified states (which is 3 in our case), $N(S_i, S_j)$ represents the number of observations in a sample which shift from state S_i to state S_j. As presented in the figure of state transition frequency pattern based on the given training sample (Figure 5), driving risk states have a high tendency to stay in the same state for a short time period, while more state shifts are observed for higher risk states as time increases.



	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0.22	0	0
4	0	0	0.11	0.22	0.11
5	0	0	0	0.22	0.11

Figure 4: Calculation of risk level cooccurrence matrix.

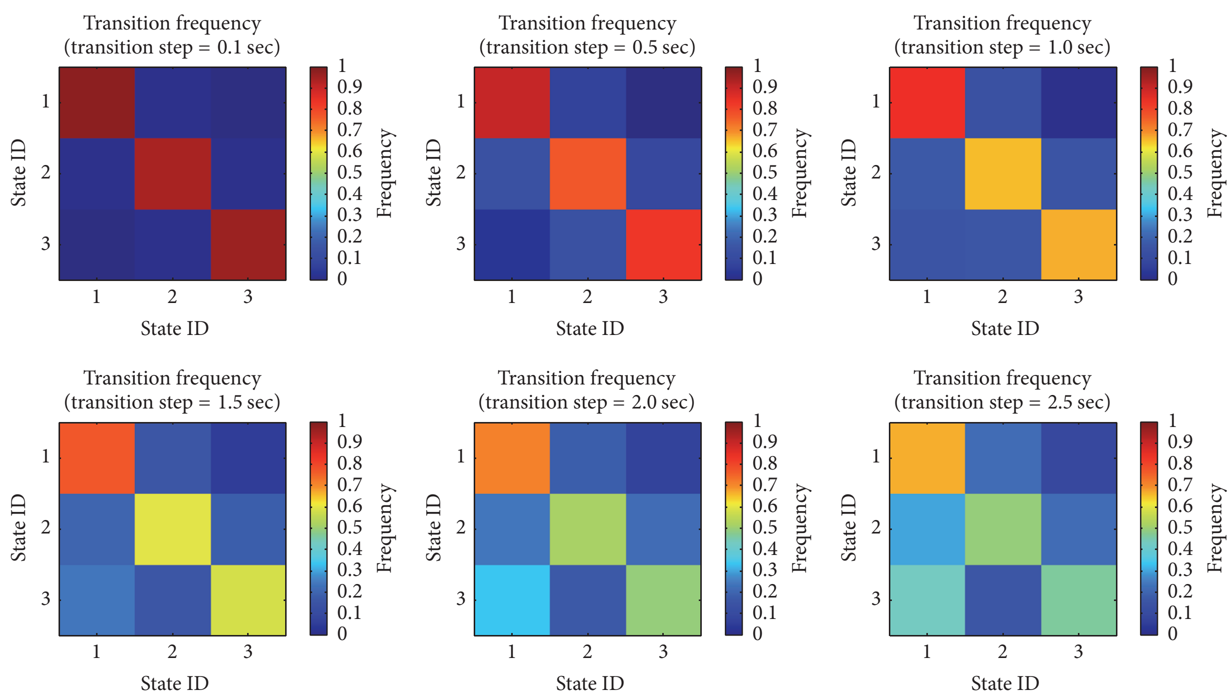


Figure 5: Driving risk state transition pattern (TW = 1.0 s).

In order to improve prediction accuracy, the paper utilized the Multinomial Logistic (MNL) regression model to characterize the pattern of transitions between states. The MNL-based category selection method has good stability and has been widely used for prediction in engineering [22]. In line with the definition of the rolling time window in Section 4.1, the probability of the time window being in state S_j at time $t+1$ given the current observed time window (at time t) being in state S_i could be estimated via the following MNL formula:

$$a_{ij}(\mathbf{Z}_t) = \Pr\left(q_{t+1} = S_j \mid q_t = S_i\right) = \frac{e^{\boldsymbol{\beta}_j^{(i)} \cdot \mathbf{Z}_t^{(i)}}}{\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k^{(i)} \cdot \mathbf{Z}_t^{(i)}}}, \quad i, j = 1, 2, 3, \tag{7}$$

where $\mathbf{Z}_t = [z_{1t}, z_{2t}, \ldots, z_{mt}]$ refers to the independent variable vector, which consists of the time window feature variables $\mathrm{RL}_{\mathrm{avg}}$, $\mathrm{RL}_{\mathrm{last}}$, CON (obtained from Level 1 data as listed in Table 1) as well as the driver, road, and environment information variables (Level 2 data). $\boldsymbol{\beta}_j$ represents the regression coefficient vector for the jth risk category in MNL model. The index value i within the parentheses in the upper right of $\mathbf{Z}_t$ and $\boldsymbol{\beta}_j$ means the estimated model conditioning on observations being in state S_i. Accordingly, $K = 3$ datasets need to be generated conditional on the state of the previous time window, resulting in $K = 3$ MNL models estimating transition probabilities for each of the three risk states separately. Considering the relative small number of observations in state shifting (i.e., $q_{t+1} \neq q_t$) given the sample (as shown in Figure 5), in order to improve

Initialization:

(1) Calculate and obtain the feature variable vector $\mathbf{x}_0 = [\mathrm{RL}_{\mathrm{avg}0}, \mathrm{RL}_{\mathrm{last}0}, \mathrm{CON}_0]$ for time window at $t = 0$.

(2) Calculate the initial state probability $\boldsymbol{\pi}_0$ by Eqn. (5)

(3) Obtain the corresponding independent variable vector for the initially observed time window:
$\mathbf{Z}_0 = [\mathrm{RL}_{\mathrm{avg}0}, \mathrm{RL}_{\mathrm{last}0}, \mathrm{CON}_0, \mathrm{DM}_0] = [\mathbf{x}_0, \mathrm{DM}_0]$

MNL-based state transition probability estimation:

(4) For $n = 1$

(4.1) Calculate $A(\mathbf{Z}_0)$ by Eqn. (7).

(4.2) Calculate the state probability distribution at $t = n = 1$ by Markov property: $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_0 A(\mathbf{Z}_0) = [\pi_{1S_1}, \pi_{1S_2}, \pi_{1S_3}]$

(4.3) Estimate the three-dimension feature variable vector $\mathbf{x}_1$by solving a set of three equations $\{\pi_{1S_i} = (1/\rho(\mathbf{x}_1, \mathbf{c}_i))/(\sum_{j=1}^{3} 1/\rho(\mathbf{x}_1, \mathbf{c}_j))\}$, $i = 1, 2, 3$ according to Eqn. (5)

(4.4) Obtain the updated independent variable vector $\widehat{\mathbf{Z}}_1 = [\widehat{\mathbf{x}}_1, \mathrm{DM}_0]$ (assuming the driving mode remains unchanged).

(5) For $n = 2, \ldots, T-1$

(5.1) Calculate $A(\widehat{\mathbf{Z}}_{n-1})$ by Eqn. (7).

(5.2) Calculate the state probability distribution at $t = n$ by Markov property: $\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n-1} A(\widehat{\mathbf{Z}}_{n-1}) = [\pi_{nS_1}, \pi_{nS_2}, \pi_{nS_3}]$

(5.3) Estimate the three-dimension feature variable vector $\mathbf{x}_n$in the same way as step (4.3).

(5.4) Obtain the updated independent variable vector $\widehat{\mathbf{Z}}_n = [\widehat{\mathbf{x}}_n, \mathrm{DM}_0]$ (same assumption as step (4.4)).

Outputs:

(6) Return the predicted state probability distribution for time window $t = T$:
$\boldsymbol{\pi}_T = \boldsymbol{\pi}_0 A(\mathbf{Z}_0) A(\widehat{\mathbf{Z}}_1) \cdots A(\widehat{\mathbf{Z}}_{T-1})$

Algorithm 1: The description of MNL-based Markov chain risk state prediction algorithm.

the convergence performance of the MNL models, instead of using the whole high-dimensional set of driver, road, and environmental information variables (Level 2 variables), a new "driving mode" ("DM") variable was constructed in the paper by clustering these Level 2 variables (could also be regarded as driving pattern characteristics) into K' patterns by elbow method. MNL model training results would be discussed in Section 6. In sum, the independent variable vector for MNL models could be expressed as $\mathbf{Z}_t = [\mathrm{RL}_{\mathrm{avg}t}, \mathrm{RL}_{\mathrm{last}t}, \mathrm{CON}_t, \mathrm{DM}_t]$.

According to the Markov property, starting from the current observed time window at $t = 0$, the risk state of any future time window at $t = T$ (i.e., to proceed with $\Delta = T$ transition steps in a Markov chain) could be determined by the initial state distribution probability $\boldsymbol{\pi}_0$ and T sets of MNL-based one-step state transition probability matrix (estimated in (7)) as follows:

$$\begin{aligned} \boldsymbol{\pi}_T &= \boldsymbol{\pi}_{T-1} A\left(\mathbf{Z}_{T-1}\right) = \boldsymbol{\pi}_{T-2} A\left(\mathbf{Z}_{T-2}\right) A\left(\mathbf{Z}_{T-1}\right) = \cdots \\ &= \boldsymbol{\pi}_0 A\left(\mathbf{Z}_0\right) A\left(\mathbf{Z}_1\right) \cdots A\left(\mathbf{Z}_{T-1}\right). \end{aligned} \tag{8}$$

In the actual prediction problem, when the number of transition steps $\Delta \geq 2$ (i.e., to predict future risk state for time window at $t = 2$ and its subsequent time windows given the observed independent variable vector at $t = 0$), the independent variable vectors $\mathbf{Z}_t$, $t = 1, \ldots, T-1$ (for time window$t = 1$ and its subsequent time windows), cannot be observed and cannot be directly applied either to (7) for estimating state transition probability or to (8) for predicting the state of the target time window. In order to meet the requirement for prediction and to improve the prediction accuracy, a MNL-based Markov chain algorithm with recursive feature variable estimation (referred to as RMNL-Markov) was proposed, as described in Algorithm 1. The key idea of the proposed algorithm is that the state probability distribution could be determined by both of the Euclidean distance (to risk state cluster centroids) based estimation method (see (5)) and Markov property (see (8)), which leads to a set of three equations solving the three-dimension future feature variable $\mathbf{x}_t = [\mathrm{RL}_{\mathrm{avg}t}, \mathrm{RL}_{\mathrm{last}t}, \mathrm{CON}_t]$, $t = 1, \ldots, T-1$.

In order to validate the proposed algorithm, two baseline algorithms (as listed below) would also be performed and compared using the same training set (with results discussed in the next section):

(1) Freq-Markov: frequency-based state transition probability estimation algorithm (see (6))

(2) CMNL-Markov: MNL-based state transition probability estimation with constant $\mathbf{Z}_t$ (assuming that the future independent variable vectors stay unchanged over time: $\mathbf{Z}_t = \mathbf{Z}_0$, $t = 1, \ldots, T-1$)

5. Model Training

The training of the model mainly includes optimal selection of three parameters: (1) the length of time window w, (2) the length of transition step δ, and (3) the number of predicted transition steps Δ. Among them, the length of time window w would directly affect the classification of risk state, the length of transition step δ would affect the estimation of state transition probability, and the number of predicted transition steps Δ would affect the prediction length (vision) of the model. The optimal combination of the parameters $\{w, \delta, \Delta\}$ would be searched via grid searching and discussed in the following sections.

Table 3: Classification of prediction results.

Observed	Predicted	
	Safe (low & medium risk state)	Dangerous (high risk state)
Safe (low & medium risk state)	True negative (TN)	False positive (FP)
Dangerous (high risk state)	False negative (FN)	True positive (TP)

5.1. Selected Evaluation Indexes. The selection of evaluation index of model performance establishes the basis for determining optimal parameters. Taking into account the possible differences in prediction accuracy across different risk states, overall and state-based prediction accuracy measurements were both employed to evaluate the prediction performance of the model.

5.1.1. Overall Prediction Accuracy Measurement. An overall prediction accuracy rate was calculated using the following formula:

$$R_{\text{acc}} = \frac{1}{n}\sum_{j=1}^{n}\delta\left(s'_j, s_j\right)$$
$$\delta\left(s'_j, s_j\right) = \begin{cases} 1, & s'_j = s_j \\ 0, & s'_j \neq s_j, \end{cases} \tag{9}$$

where $\delta(s'_j, s_j)$ represents a Dirac delta function, s'_j and s_j are the predicted and the observed risk states, respectively, and n is the total number of predictions made. The overall prediction accuracy rate measures the overall prediction performance in accuracy for all risk states.

5.1.2. State-Based Prediction Accuracy Measurement. As most states tend to stay unchanged over small-to-medium-sized transition steps (as presented in Figure 5), the performance of predicting state-shifting cases (i.e., $q_{t+1} \neq q_t$ in (7)) should be evaluated separately to eliminate the dominant effects of prediction performance for state-staying cases. As a result, general and state-shifting prediction accuracy rates were utilized for state-based prediction performance evaluation. Such state-shifting measurement could be critical for real-world application as a high prediction accuracy rate is usually required for forecasting changing state scenarios, especially for those shifting from lower risk states to higher risk states.

The state-based general prediction accuracy rate could be determined as follows:

$$R_{\text{acc}}{}^{(i)} = \frac{1}{n^{(i)}}\sum_{j=1}^{n^{(i)}}\delta\left(s'_j, s_j{}^{(i)}\right), \quad i = 1, 2, 3, \tag{10}$$

where $s_j{}^{(i)}$ represents the observed ith risk state and $n^{(i)}$ is the total number of observations in the ith risk state. The general state-based prediction accuracy rate measures the overall prediction accuracy for each of the risk states.

The state-based state-shifting (SS) prediction accuracy rate was calculated as follows:

$$R_{\text{accSS}}{}^{(i)} = \frac{1}{n_{\text{SS}}^{(i)}}\sum_{j=1}^{n_{\text{SS}}^{(i)}}\delta\left(s'_j, s_j{}^{(i)}\right), \quad i = 1, 2, \ldots, K, \tag{11}$$

where $n_{\text{SS}}^{(i)}$ is the total number of observations in the ith risk state which are shifted from the jth ($j \neq i$) state. The accuracy rate of state-shifting prediction measures the prediction accuracy for each of the risk states that involve state shifting.

5.1.3. Additional Prediction Measurement. For two-category risk classification problem (e.g., being safe versus dangerous), true positive rate (TPR) and false positive rate (FPR) are usually employed to assess prediction performance from the perspective of users [23]. In the paper, the low and medium risk states were combined into a "safe" category to perform the TPR and FPR calculation, as shown in Table 3.

TPR and FPR are defined as follows:

$$\text{TPR} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$
$$\text{FPR} = \frac{\text{FP}}{(\text{FP} + \text{TN})}. \tag{12}$$

As could be noted from equations above, a higher TPR (the rate of observed high risk states correctly predicted to be high risk states) indicates a higher percentage of correctness in predicting the high risk states, and accordingly the prediction system is believed to be more effective in preventing accidents. However, seeking a high TPR may also cause overfitting problem of the model which should be avoided, and thus it is necessary to ensure that the FPR (the rate of observed medium and low risk states wrongly predicted to be high risk states) from the prediction model lies within a reasonable range. Considering the tolerance of drivers for false warnings, 5% is usually chosen as the highest FPR in practice.

The following sections would present and discuss the selection of parameters based on the selected prediction performance evaluation indexes.

5.2. Parameter Selection and Overall Prediction Accuracy. Considering the accuracy and timelines requirement for driving risk state prediction, the length of time window $w \in [0.1, 2.0]$ sec (gridded at 0.1 sec), the length of transition step $\delta \in [0.1, 2.0]$ sec (gridded at 0.1 sec), and the number of predicted transition steps $\Delta \in [1, 10]$ (gridded at 1) were

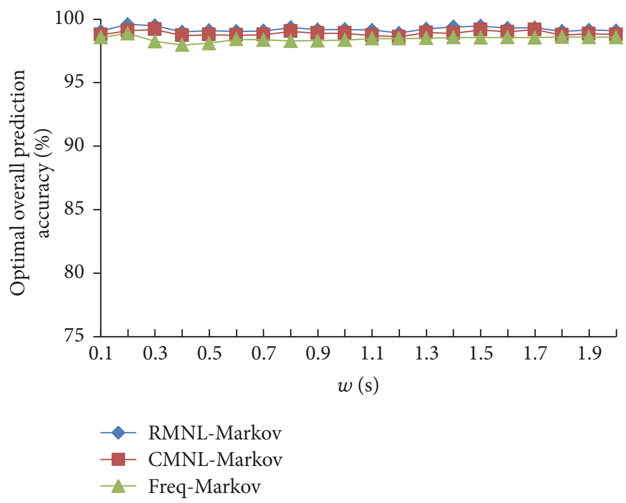


FIGURE 6: Prediction accuracy and time window length.

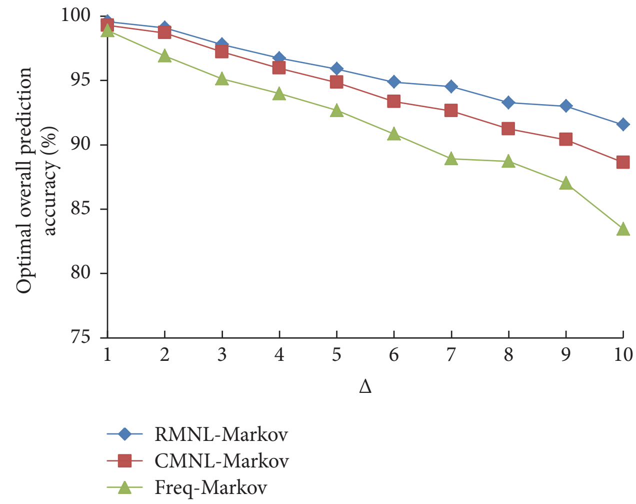


FIGURE 8: Prediction accuracy and number of transition steps.

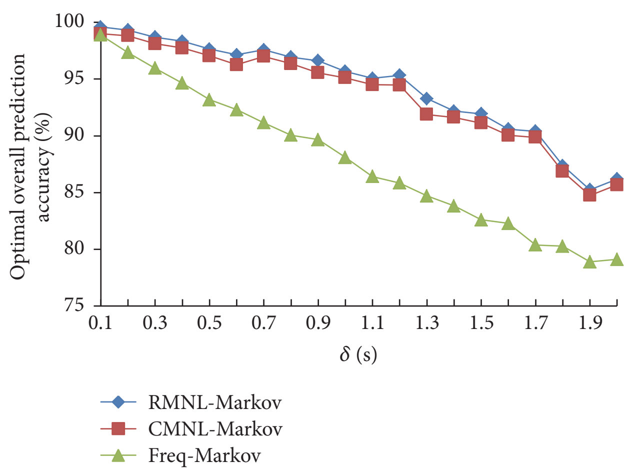


FIGURE 7: Prediction accuracy and transition step length.

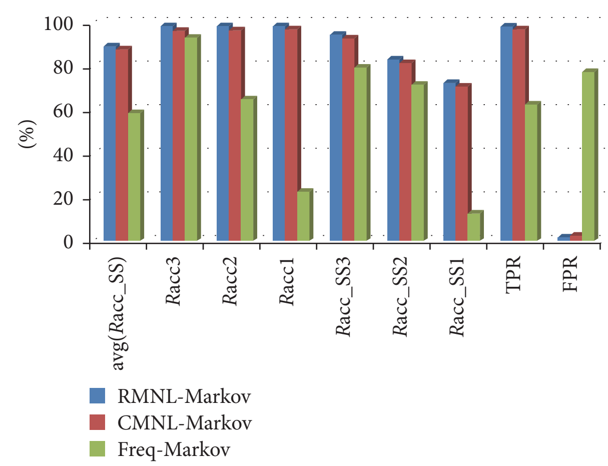


FIGURE 9: Mean prediction results of the top 10 parameter combinations.

experimented via grid searching for optimal prediction outcome. Among all the prediction results from these parameter value combinations, the highest overall prediction accuracy rate under a given time window length was taken as its optimal overall prediction accuracy rate. Figure 6 presents the trend of optimal overall prediction accuracy rates over different time window lengths. In the same way, the trend of optimal overall prediction accuracy rates over different transition step lengths and number of predicted transition steps are shown in Figures 7 and 8.

Figures 6–8 show that the optimal overall prediction accuracy rate does not increase or decrease significantly with the change of the time window length w, while with the increase in the transition step length δ and the number of predicted transition steps Δ, the optimal overall prediction accuracy rate decreases gradually. The reason for the decrease in accuracy may be that as δ and Δ increase, the length of prediction (vision to future) also increases, which makes the prediction more challenging given the same amount of current state information. Also, results show that the two MNL-based models perform much better than the frequency-based model when δ increases (Figure 7), validating the effectiveness of such MNL-based state transition estimation methodology. In addition, compared with the CNML-Markov model, the increase in number of transition steps causes less reduction in the optimal overall prediction accuracy rate for the RMNL-Markov model (Figure 8), indicating the effectiveness of the proposed recursive feature variable estimation method.

5.3. Parameter Selection and State-Based Prediction Accuracy, TPR, and FPR. Considering the high accuracy requirement for predicting state-shifting (SS) cases in real-time driving risk prediction application, the average accuracy rates in predicting SS risk states under different combinations of parameter values were sorted, with the mean prediction results of the highest 10 combinations listed in Figure 9. It could be noted that the state-based general prediction

Table 4: Ten best parameter combinations for RMNL-Markov model.

Comb. number	1	2	3	4	5	6	7	8	9	10
w	1.9	1.9	1.7	1.7	1.7	1.4	1.4	1.9	1.7	1.4
δ	0.2	0.2	0.1	0.2	0.1	0.4	0.4	0.9	0.9	0.7
Δ	1	2	1	1	2	1	2	1	1	1

accuracy rates $R_{acc}{}^{(i)}$, $i = 1, 2, 3$ (labeled as Racc1, Racc2, and Racc3 in Figure 9), are generally higher than the state-based SS prediction accuracy rates $R_{acc_{SS}}^{(i)}$, $i = 1, 2, 3$ (labeled as Racc_SS1, Racc_SS2, and Racc_SS3 in Figure 9), which may be due to the high ratio of the number of observed state retentions to that of observed state shifts in the sample (Figure 5). Consistent with the prediction performance shown in Figures 6–8, the proposed RMNL-Markov model also maintains the highest state-based prediction accuracy rates compared to the two baseline models and thus was selected for the following analysis.

The 10 parameter combinations with the top 10 average prediction accuracy rates in predicting SS risk states using the RMNL-Markov model are listed in Table 4, with prediction results presented in Figure 10. It could be noted that all the average SS prediction accuracy rates of the top 10 combinations are more than 80%, and all of them could meet the 5% FPR level. Also, the first 7 combinations have higher SS prediction for higher risk states than that for lower risk states, which could better meet the practical early warning requirements on timely prediction of higher risk states. Among them, the number 7 combination features a longer prediction length ($= \delta \times \Delta = 0.4 \times 2 = 0.8$ sec) while maintaining a relatively high prediction accuracy (its SS prediction accuracy rates are more than 85% for both high risk and medium risk states). Taking into account both the accuracy and timeliness requirements for driving risk prediction model, parameter combination number 7 was selected as the final optimal parameters for the proposed RMNL-Markov model.

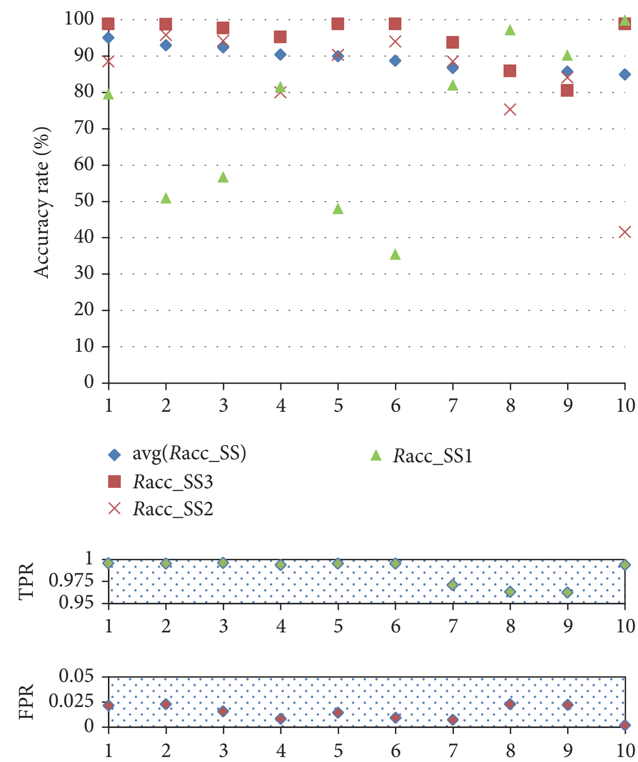


Figure 10: Prediction results of the top 10 parameter combinations for RMNL-Markov.

5.4. Prediction Model with Selected Parameters

5.4.1. Classified Driving Risk States. Based on the selected time window length parameter $w = 1.4$ sec, a total of 10,651 rolling time windows were obtained from the training sample. The clustering results of the time window-based risk states are summarized in Table 5 and Figure 11.

As presented in Table 5, with the risk state getting higher, its corresponding average iTTC has an increasing trend, while its average THW has a decreasing trend, which is consistent with what one would expect. At the same time, two points should be noted from Figure 11: (1) in Figure 11(c), some observations with an average iTTC $\in (0, 0.4]$ (i.e., TTC > 2.5 sec) but with low average THW (<2 sec) and positive CON (indicating an increasing trend in the risk level sequence within the time window) are allocated to the high risk state; (2) in Figure 11(b), some observations with an average iTTC > 1 (i.e., TTC < 1 sec) but with negative CON (indicating an decreasing trend in the risk level sequence within the time window) and ending in low risk level (at the last time point of the time window) are allocated to the medium risk state rather than the high risk state. Such parameter value discrepancies exist between the classified risk state here and the traditional TTC-based risk classification because the risk states in the paper were defined based on the TTC-and-THW two-dimension plane, of which prediction values are not comparable to the single parameter prediction results by other approaches in the literature. Anyhow, it could be noted that the high risk cluster has an average TTC value around 2.2 sec (iTTC_avg $= 0.453$ sec in Table 5) which is close to the 2.4 sec TTC warning threshold recommended by NHTSA [24], which to some extent validates the risk state classification results here.

5.4.2. Estimated State Transition Probability. Given the training sample and based on the selected parameters $\{w = 1.4\text{ sec}, \delta = 0.4\text{ sec}\}$, a total of 2,643 independent-dependent variable pairs were obtained for MNL training based on the selected parameter combination, with training results shown in Table 6.

Results show that the driving mode variable has a significant impact on state transition probability for each risk state (at $\alpha = 0.05$ level). The results prove the necessity of including the real-time driver, road, and environment characteristics in developing the Markov chain model for driving risk state prediction.

Table 5: Characteristics of TW-based driving risk state clusters.

(TW = 1.4 s)	Cluster group	C1 (low risk)	C2 (medium risk)	C3 (high risk)
Clustering variables	RL_avg	2.329	5.027	7.115
	RL_last	2.293	5.053	7.484
	CON	−0.054	−0.002	0.188
Observed parameters	iTTC_avg	−0.065	0.114	0.453
	iTTC_std	0.029	0.039	0.226
	THW_avg	2.471	1.254	0.666
	THW_std	0.239	0.059	0.067
	# of obs.	5587 (52.4%)	3064 (28.8%)	2000 (18.8%)

Table 6: MNL training results based on test set.

Variables	Future state: 1	Future state: 2	Future state: 3
Initial state: (1) low risk			
Driving mode	-0.474^{*}	-0.497^{**}	ref.
RL_avg	-0.569^{**}	-0.151^{*}	ref.
RL_last	-0.150^{*}	0.543^{**}	ref.
CON	0.153^{*}	-0.213^{*}	ref.
Constant	9.608	6.194	ref.
$L(0)$		−321.583	
$L(B)$		−128.311	
Rho-square		0.601	
Initial state: (2) medium risk			
Driving mode	-0.105^{*}	-0.018^{*}	ref.
RL_avg	0.857^{**}	0.279^{*}	ref.
RL_last	1.587^{**}	0.599^{*}	ref.
CON	-0.672^{*}	-0.172^{*}	ref.
Constant	−0.421	2.483	ref.
$L(0)$		−144.884	
$L(B)$		−26.989	
Rho-square		0.814	
Initial state: (3) high risk			
Driving mode	0.038	0.096^{*}	ref.
RL_avg	-0.025^{*}	-0.052^{*}	ref.
RL_last	-0.264^{*}	-0.619^{**}	ref.
CON	0.054	0.125^{*}	ref.
Constant	−27.884	−2.213	ref.
$L(0)$		−161.964	
$L(B)$		−19.525	
Rho-square		0.880	

Note. "ref." represents the reference state in MNL regression. *Statistical t-test significance, $p < 0.05$. **Statistical t-test significance, $p < 0.01$. "$L(0)$" represents initial likelihood of the model with constants only; "$L(B)$" represents likelihood of the specified model. Pseudo-$R^2 = 1 - L(0)/L(B)$.

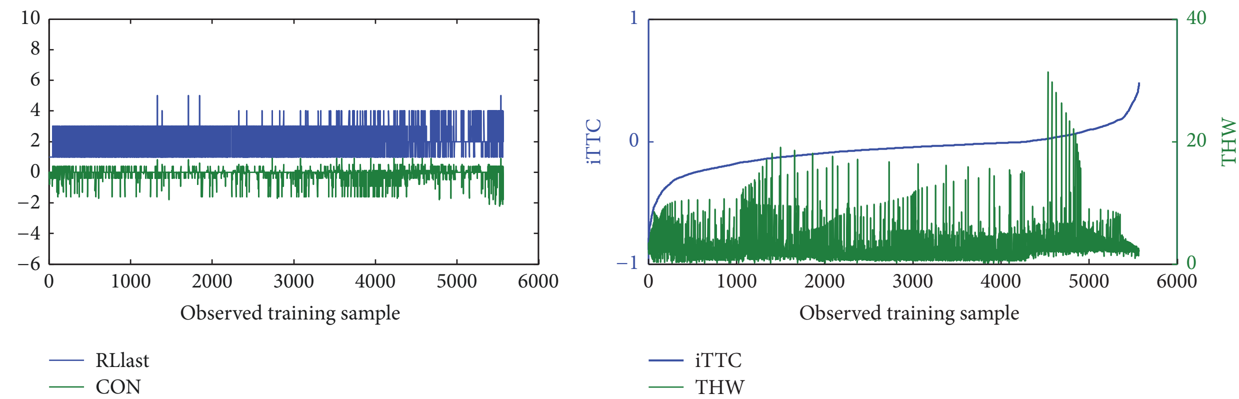
6. Model Validation

Given the selected parameters $\{w = 1.4\,\text{sec}, \delta = 0.4\,\text{sec}, \Delta = 2\}$, a total of 877 observation-prediction time window pairs were obtained based on the test set for Markov chain forecasting using the established RMNL-Markov model in Section 5, of which 172 risk state shifts were observed. TPR and FPR of the model prediction are 96.6% and 2.7%, respectively, indicating that the established model has a good overall prediction performance in accuracy. Results of state-shifting (SS) prediction (in Table 7) show that the average SS prediction accuracy rate is 85.3% across the states, with the highest SS prediction accuracy rate at 90.0% for high risk state, indicating that the established model could effectively predict higher risk driving conditions.

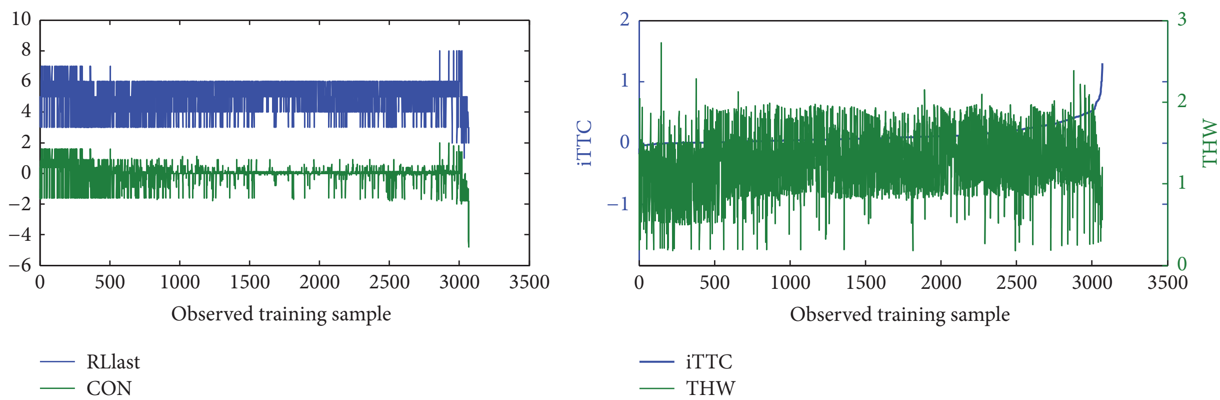
The time-series samples in the testing set were also used as virtual online observations (as virtual online tests) to validate the timeliness performance of the proposed model. The moment when a high risk state was correctly predicted was recorded and compared to the moment when the high risk state was first observed in real time. Results show that on average the model could make correct high risk state predictions 0.7 sec earlier than the real high risk state occurs. As previous research shows, at least 60%~90% of rear-end accidents could be avoided as long as the driver could be warned 0.5~1.0 sec prior to the collision risk [25]; the established model is promising for early warning in reducing most of the accidents.

7. Conclusion

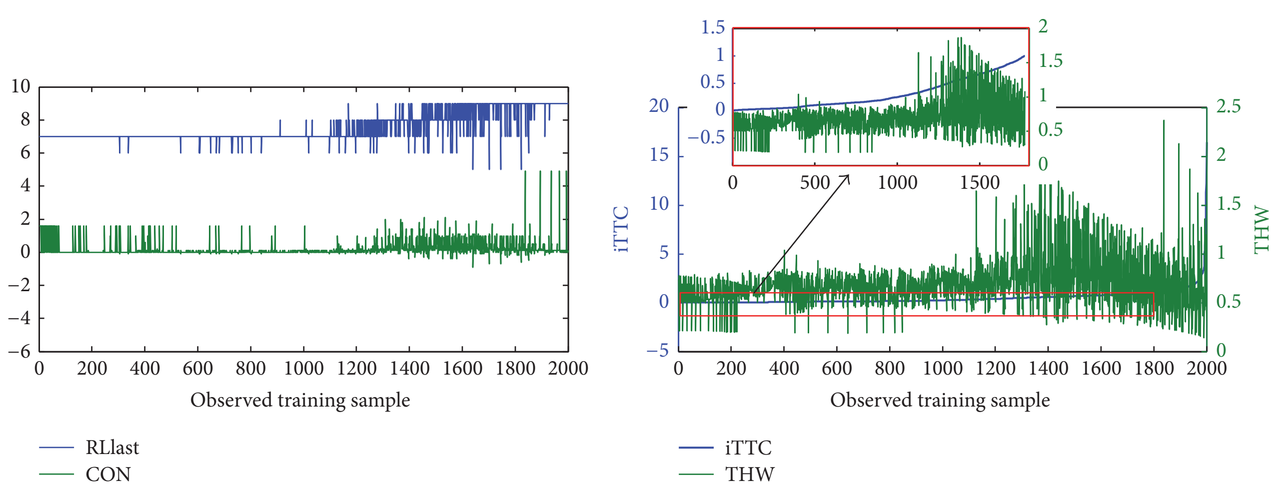
A MNL-based Markov chain model was proposed for driving risk state prediction. The prediction model was trained and validated using the "100-car" NDS data from Virginia Tech. The results show that, compared to the traditional frequency-based state transition probability estimation method, the recursive MNL-based algorithm proposed in the paper could capture the comprehensive effects of driver, road, and environment on the evolution of driving risk states and obtain promising prediction results. Prediction accuracy rate for states shifting to medium and high risk states could reach over 85% under 5% FPR, and the virtual online tests show that the proposed algorithm could generally meet the timely requirement of early warning for collision avoidance. More future works would be focused on model improvement and verification by collecting more driving risk observations through NDS and conducting vehicle tests to further validate the real-time prediction performance of the model. Note that the proposed recursive MNL-based Markov chain model could also be applied to other domains featuring state evolution process such as real-time traffic state prediction for traffic management and fault state prediction for predictive



(a) Cluster 1



(b) Cluster 2



(c) Cluster 3

FIGURE 11: Driving risk level time window clustering results. *Note*. The random observed time window samples were sorted according to the average value of iTTC within the time window for easier understanding.

TABLE 7: SS prediction results based on test set.

		Predicted SS state			
	Risk state	*S*1Low	*S*2Medium	*S*3High	Accuracy rate
Observed SS state	*S*1	63	11	4	80.80%
	*S*2	3	46	5	85.20%
	*S*3	1	3	36	90.00%
Average					85.30%

maintenance. Thus, such a study not only provides new basis for driving safety evaluation but also offers significant potential for engineering applications.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Projects nos. 61773184 and 61573171).

References

[1] World Health Organization, 2016, http://www.who.int/mediacentre/factsheets/fs358/en/.

[2] H. Jula, E. B. Kosmatopoulos, and P. A. Ioannou, "Collision avoidance analysis for lane changing and merging," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 6, pp. 2295–2308, 1999.

[3] H. Zhang, S. Qiu, and H. Li, "Evaluation of Risk Judgement Indices for Rear-end Collisions," in *Proceedings of the in Proceedings of the 2011 Symposium on automotive safety technology*, August 2011.

[4] C. Wang, R. Fu, Q. Zhang et al., "Research on Parameter TTC Characteristics of Lane Change Warning System," *China Journal of Highway and Transport*, vol. 28, no. 8, pp. 91–107, 2015.

[5] J. Ni, Z. Liu, X. Tu et al., "Safety Prediction Model of Lane Changing Based on Driver Assistance System," *Journal of Transportation Systems Engineering and Information Technology*, vol. 16, no. 4, pp. 95–100, 2016.

[6] X. Xiong, L. Chen, and J. Liang, "A new framework of vehicle collision prediction by combining SVM and HMM," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12.

[7] J. Wang, J. Wu, X. Zheng, D. Ni, and K. Li, "Driving safety field theory modeling and its application in pre-collision warning system," *Transportation Research Part C: Emerging Technologies*, vol. 72, pp. 306–324, 2016.

[8] G. Zhang, K. K. W. Yau, X. Zhang, and Y. Li, "Traffic accidents involving fatigue driving and their extent of casualties," *Accident Analysis & Prevention*, vol. 87, pp. 34–42, 2016.

[9] S. Klauer, "The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data," DOT-HS-810-594, US Department of Transportation, Washington, WA, USA, 2006.

[10] R. Talbot, H. Fagerlind, and A. Morris, "Exploring inattention and distraction in the SafetyNet Accident Causation Database," *Accident Analysis & Prevention*, vol. 60, pp. 445–455, 2013.

[11] X. Zhang, S. Wang et al., "A Research on Driving Condition Prediction for HEVs Based on Markov Chain," *Automotive Engineering*, vol. 36, no. 10, pp. 1216–1220, 2014.

[12] Z. Ma, H. N. Koutsopoulos, L. Ferreira, and M. Mesbah, "Estimation of trip travel time distribution using a generalized Markov chain approach," *Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 1–21, 2017.

[13] M. Ramezani and N. Geroliminis, "On the estimation of arterial route travel time distribution with Markov chains," *Transportation Research Part B: Methodological*, vol. 46, no. 10, pp. 1576–1590, 2012.

[14] Virginia Tech Transportation Institute, "VTTI Data Warehouse," 2014, http://forums.vtti.vt.edu/index.php.

[15] D. Lu and H. Zhang, *Stochastic Process and Application*, Press of Tsinghua University, Beijing, 2012.

[16] J. Wang, Y. Zheng, X. Li et al., "Driving risk assessment using near-crash database through data mining of tree-based model," *AccidenT Analysis & Prevention*, vol. 84, pp. 54–64, 2015.

[17] Y. Zheng, J. Wang, X. Li, C. Yu, K. Kodaka, and K. Li, "Driving risk assessment using cluster analysis based on naturalistic driving data," in *Proceedings of the 2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014*, pp. 2584–2589, China, October 2014.

[18] C. Hydén, "Traffic safety work with video-processing," Tech. Rep., Transportation Department, University Kaiserslautern, 1996.

[19] D. J. Ketchen Jr. and C. L. Shook, "The application of cluster analysis in strategic management research: An analysis and critique," *Strategic Management Journal*, vol. 17, no. 6, pp. 441–458, 1996.

[20] Y. Li, J. Lu, and K. Xu, "Crash risk prediction model of lane-change behavior on approaching intersections," *Discrete Dynamics in Nature and Society*, vol. 2017, pp. 1–12, 2017.

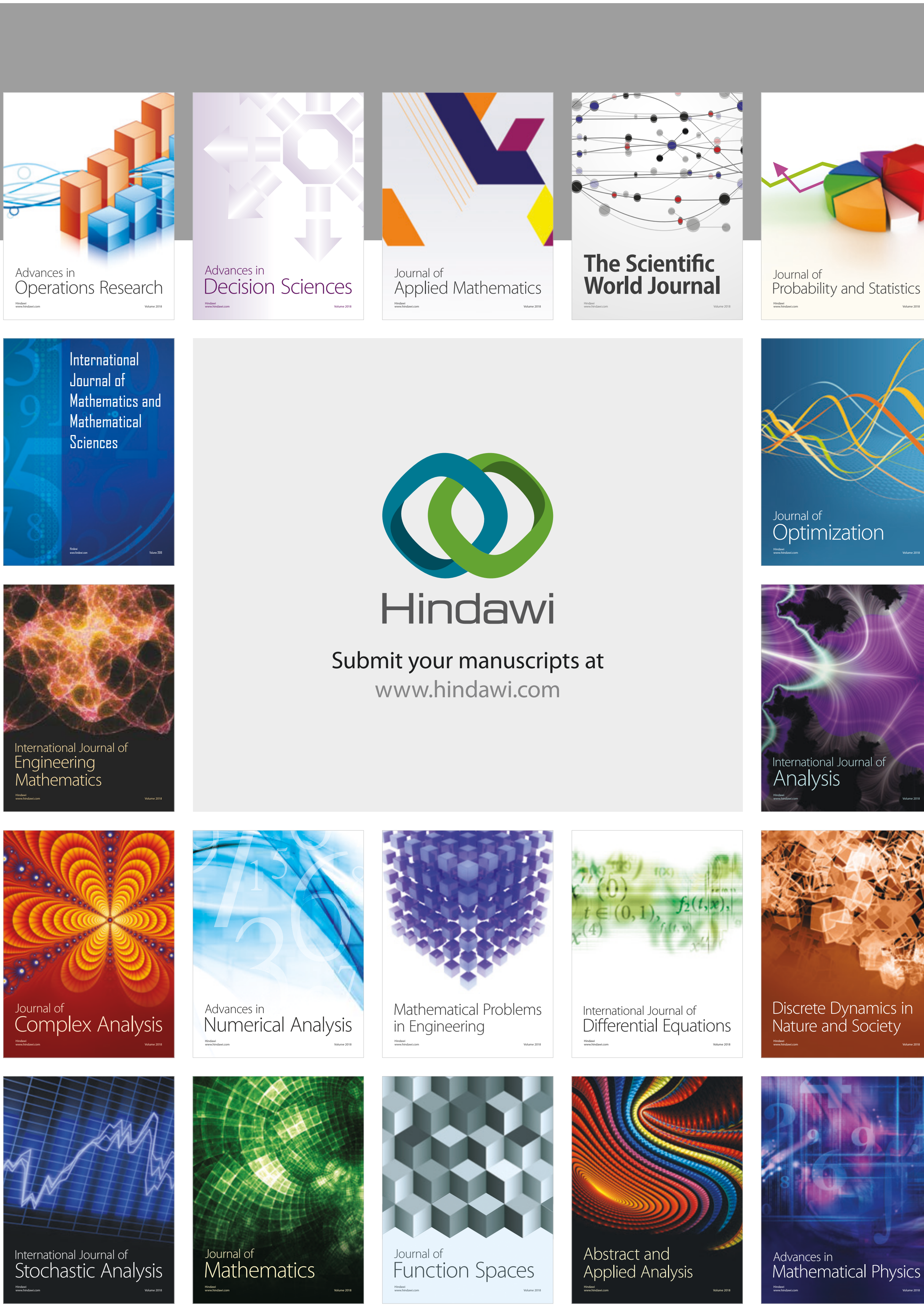
[21] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, Massachusetts, Mass, USA, 1999.

[22] J. Ledolter, *Multinomial Logistic Regression, Data Mining and Business Analytics with R*, John Wiley Sons, Inc, New Jersey, NJ, USA, 2013.

[23] G. S. Aoude, V. R. Desaraju, L. H. Stephens, and J. P. How, "Driver behavior classification at intersections and validation on large naturalistic data set," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 2, pp. 724–736, 2012.

[24] S. J. Brunson, E. M. Kyle, N. C. Phamdo, and G. R. Preziotti, "Alert Algorithm Development Program: NHTSA Rear-end Collision Alert Algorithm," Tech. Rep., US Department of Transportation, Washington, Wash, USA, 2002.

[25] H. H. Meinel, "Automotive millimeterwave radar history and present status," in *Proceedings of the 1998 28th European Microwave Conference (EuMC '98)*, pp. 619–629, October 1998.


Advances in
Operations Research
Advances in
Decision Sciences
Journal of
Applied Mathematics
The Scientific
World Journal
Journal of
Probability and Statistics
International
Journal of
Mathematics and
Mathematical
Sciences
Hindawi
Submit your manuscripts at
www.hindawi.com
Journal of
Optimization
International Journal of
Engineering
Mathematics
International Journal of
Analysis
Journal of
Complex Analysis
Advances in
Numerical Analysis
Mathematical Problems
in Engineering
International Journal of
Differential Equations
Discrete Dynamics in
Nature and Society
International Journal of
Stochastic Analysis
Journal of
Mathematics
Journal of
Function Spaces
Abstract and
Applied Analysis
Advances in
Mathematical Physics