Accident Analysis and Prevention 117 (2018) 250–261

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

How to determine an optimal threshold to classify real-time crash-prone traffic conditions?

Kui Yang[a,b], Rongjie Yu[a,b,*], Xuesong Wang[a,b], Mohammed Quddus[c], Lifang Xue[d]

[a] School of Transportation Engineering, Tongji University, 4800 Cao'an Road, 201804, Shanghai, China
[b] The Key Laboratory of Road and Traffic Engineering, Ministry of Education, 4800 Cao'an Road, 201804, Shanghai, China
[c] School of Civil and Building Engineering, Loughborough University, Loughborough LE11, 3TU, United Kingdom
[d] College of Arts and Sciences, Shanxi Agricultural University, No. 1 Mingxian Nan Road, 030801, Taigu, Shanxi, China

ARTICLE INFO

Keywords:
Urban expressway safety management
Crash risk evaluation
Mixed logit model
Threshold selection method
Cross-entropy
Between-class variance

ABSTRACT

One of the proactive approaches in reducing traffic crashes is to identify hazardous traffic conditions that may lead to a traffic crash, known as real-time crash prediction. Threshold selection is one of the essential steps of real-time crash prediction. And it provides the cut-off point for the posterior probability which is used to separate potential crash warnings against normal traffic conditions, after the outcome of the probability of a crash occurring given a specific traffic condition on the basis of crash risk evaluation models. There is however a dearth of research that focuses on how to effectively determine an optimal threshold. And only when discussing the predictive performance of the models, a few studies utilized subjective methods to choose the threshold. The subjective methods cannot automatically identify the optimal thresholds in different traffic and weather conditions in real application. Thus, a theoretical method to select the threshold value is necessary for the sake of avoiding subjective judgments. The purpose of this study is to provide a theoretical method for automatically identifying the optimal threshold. Considering the random effects of variable factors across all roadway segments, the mixed logit model was utilized to develop the crash risk evaluation model and further evaluate the crash risk. Cross-entropy, between-class variance and other theories were employed and investigated to empirically identify the optimal threshold. And K-fold cross-validation was used to validate the performance of proposed threshold selection methods with the help of several evaluation criteria. The results indicate that (i) the mixed logit model can obtain a good performance; (ii) the classification performance of the threshold selected by the minimum cross-entropy method outperforms the other methods according to the criteria. This method can be well-behaved to automatically identify thresholds in crash prediction, by minimizing the cross entropy between the original dataset with continuous probability of a crash occurring and the binarized dataset after using the thresholds to separate potential crash warnings against normal traffic conditions.

1. Introduction

Given the technological progress over the last decade in traffic data detection, storage and mining, real-time crash prediction has become a popular research topic within the safety community. Crash risk evaluation and threshold selection are the two essential steps of real-time crash prediction. Crash risk evaluation is related to the investigation of a relationship between the crash occurrence and geometric characteristics, real-time traffic flow parameters, weather conditions. The relationship is used to evaluate the probability of a crash occurring given a specific traffic condition and identify hazardous traffic conditions. The threshold selection is to investigate the algorithm to select the optimal cut-off point of the posterior probability (0 < posterior probability < 1, also known as crash risk), which is used for separating potential crash warnings from normal traffic conditions, and further triggering Active Traffic Management (ATM) control strategies (Abdel-Aty et al., 2006). Previous studies mostly focused on crash risk evaluation, with the purpose of identifying impact factors on crash occurrence in order to further understand the crash mechanisms (e.g. Abdel-Aty et al., 2005; Yu and Abdel-Aty, 2013a), modeling technique aimed at better classification accuracy (e.g. Abdel-Aty and Pande, 2005; Yu and Abdel-Aty, 2013b; Xu et al., 2013a). However, there is a dearth of study focused on how to determine a reliable threshold for real-time crash prediction.

In real-time crash prediction and its application in ATM, an appropriate threshold should be used to compare with the estimated

* Corresponding author at: School of Transportation Engineering, Tongji University, 4800 Cao'an Road, 201804, Shanghai, China.
E-mail address: yurongjie@tongji.edu.cn (R. Yu).

https://doi.org/10.1016/j.aap.2018.04.022
Received 14 September 2017; Received in revised form 20 March 2018; Accepted 19 April 2018

probability which is the output of the crash risk evaluation model on the basis of real-time traffic data. If the probability exceeds the predetermined threshold, the case is predicted as a *potential crash scenario*, and then a crash warning is alerted, and further control strategies are triggered (Abdel-Aty et al., 2010). In addition, there is a dilemma in selecting the correct threshold (which is a metric ranging from 0 to 1) because a high threshold normally fails to identify many potential crash conditions whereas a low threshold falsely identifies normal traffic conditions as 'hazardous'. False alarms may affect the driver's compliance and raise the cost of ATM operations.

Only a few existing studies on real-time crash prediction models involved identifying a threshold when discussing the predictive performance of the models, and their methods are subjective. Moreover, the subjective approach cannot automatically identify the optimal thresholds in different traffic and weather conditions, aimed at capturing the temporal-spatial heterogeneity of crashes which has been proved to exist by researchers (e.g. Xu et al., 2013b; Yu et al., 2016). Additionally, in order to avoid subjective judgments, a theoretical method to select the threshold value is necessary (Li and Tzeng, 2009). In other fields of pattern recognition (e.g. image segmentation, medical field), different methods of the threshold selection have been investigated for more than half a century (Zhang, 2014). Their methods promoted the progress of technology in studies and applications.

This study aims to explore available threshold selection methods so as to automatically identify the optimal threshold for real-time crash prediction. Cross-entropy, between-class variance and other theories were utilized and their performances were compared on the basis of several evaluation criteria. Traffic data and historical crash data from Shanghai Urban Expressway System were used in this analysis. Considering the random effects of variable factors across all roadway segments, mixed logit model was employed to develop the crash risk evaluation model and further evaluate the crash risk. Different thresholds were selected by five threshold selection methods, and their predictive performances were further evaluated through different evaluation criteria. Besides, 5-fold cross-validation was used to test the thresholds for deriving a more accurate estimate of prediction performance.

The rest of this paper is divided into six sections. First, previous studies on threshold selection in the real-time crash prediction and other fields are summarized. The second section describes the study area. The third section describes the data preparation procedures. Afterwards, the modeling techniques, threshold selection methods and several evaluation criteria are introduced. The fifth section presents the modelling and comparison results of different threshold selection methods. Finally, discussion and conclusions of this work are provided.

2. Literature review

2.1. Threshold selection involved in crash risk evaluation

In existing studies on real-time crash prediction, there is a dearth of research that focuses on threshold selection method. And only when discussing the prediction performance of the models, a few studies tried to choose the threshold subjectively.

When a matched case-control logistic regression model was used to develop a crash risk evaluation model, the average values of the explanatory variables associated with all non-crash cases within each matched stratum were calculated as the "normal traffic conditions", and the odds ratio of each case relative to "normal traffic conditions" within each stratum was used as the crash risk index. Thus, the odds ratio with a value equal to one was selected as the threshold (e.g. Abdel-Aty et al., 2005; Ahmed and Abdel-Aty, 2012), which is known as fixed threshold based on the odds ratio. Similar to posterior probability, the odds ratio is an index indicating the crash risk level relative to non-crashes, and thus a value greater than one is regarded subjectively as "more hazardous" than "normal traffic conditions" by fixed threshold based on the odds ratio. Moreover, the method will not work if the modeling technique is not a matched case-control logistic regression model. Therefore, fixed threshold based on the odds ratio has the uniqueness of crash risk evaluation model development technique and the fixed threshold, which creates its limitations in range of application. And it has been not accepted by all researchers.

Due to the imbalance of the proportions of crash and non-crash cases in the sample, overall classification accuracy over validation dataset would not be a good measure for model performance evaluation. Therefore, the top 30 percentile of posterior probability (i.e. first three deciles) was decided as the threshold (Pande and Abdel-Aty, 2006a, 2006b). Similarly, Pande et al. (2011) chose the top 20 percentile of posterior probability as the threshold. Since the threshold is mainly affected by the proportion of crashes in samples, different samples with the same proportion of crashes but different characteristic distributions cannot select obviously different thresholds.

Aimed at balancing the predictive accuracy of crash and non-crash, the cut-off point where the predictive accuracy of crashes was equal to that of non-crashes (i.e. the intersection of cumulative proportion curves of crash and non-crash cases), was chosen as the threshold by the intersection point method (e.g. Xu et al., 2013b). But the predictive accuracies of crashes and non-crashes among different models or strategies lack some comparability because of different weighting scores.

Totally, existing three types of threshold selection methods have no specific mathematical optimization functions. And it creates a certain amount of subjectivity and the limitations in application. Moreover, they are sensitive to a fraction of specific sample and cannot absorb all distribution information of the dataset. Thus, a theoretical method to select the threshold value is necessary so as to avoid subjective judgments.

2.2. Threshold selection in other fields

Threshold selection techniques are fundamental for image segmentation. Different techniques, ranging from a bilevel threshold selection with a single threshold to a multilevel threshold selection with multiple thresholds dividing pixels into categories, have been proposed (Yin, 2007).

The first category, known as histogram shape-based method, contains the approaches which determine the optimal threshold by analyzing the profile characteristics of the gray-level histogram of pixel in image, which is the basic information for image thresholding. With decades of experiments and applications, bimodal histogram threshold method (Weszka et al., 1974) and P-tile method (Doyle, 1962; Samopa and Asano, 2009) could achieve better binarized image, and became the two widely used shape-based methods. Bimodal histogram threshold selects the cut-off point of gray-level of pixel at valleys between two peaks as the threshold, while P-tile method requires the proportion of the object after being binarized will not be less than that in the original dataset.

The second category, known as optimization method, belongs to the techniques which determine the optimal threshold by optimizing a certain objective function or theory, which use some extra information such as spatial information or binarized images:

(i) Entropy-based methods: Maximum entropy method (Pun, 1980; Kapur et al., 1985) and minimum cross-entropy method (Yin, 2007; Li and Lee, 1993) can achieve better binarized image, and became the two widely used methods. Maximum entropy method maximizes the entropy of foreground and background regions, while the minimum cross-entropy method minimizes the cross-entropy between the original and binarized image.
(ii) Variance-based methods: the maximum between-class variance method (Otsu, 1979) selects an optimal threshold by maximizing the separability of the resultant classes in gray levels for image segmentation. And it is one of the best threshold selection methods