Accident Analysis and Prevention 50 (2013) 405–415

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

A two-stage mining framework to explore key risk conditions on one-vehicle crash severity

Yu-Chiun Chiou [a,*], Lawrence W. Lan [b], Wen-Pin Chen [a]

[a] Institute of Traffic and Transportation, National Chiao Tung University, 4F, 118, Sec. 1, Chung-Hsiao W. Rd., Taipei 100, Taiwan, ROC
[b] Department of Television & Internet Marketing Management, Ta Hwa University of Science and Technology, Institute of Traffic and Transportation, National Chiao Tung University, Taiwan, ROC

ARTICLE INFO

Article history:
Received 11 January 2012
Received in revised form 7 May 2012
Accepted 9 May 2012

Keywords:
Crash severity
Genetic mining rule
One-vehicle crashes
Mixed logit model
Stepwise rule-mining algorithm

ABSTRACT

This paper proposes a two-stage mining framework to explore the key risk conditions that may have contributed to the one-vehicle crash severity in Taiwan's freeways. In the first stage, a genetic mining rule (GMR) model is developed, using a novel stepwise rule-mining algorithm, to identify the potential risk conditions that best elucidate the one-vehicle crash severity. In the second stage, a mixed logit model is estimated, using the antecedent part of the mined-rules as explanatory variables, to test the significance of the risk conditions. A total of 5563 one-vehicle crash cases (226 fatalities, 1593 injuries and 3744 property losses) occurred in Taiwan's freeways over 2003–2007 are analyzed. The GMR model has mined 29 rules for use. By incorporating these 29 mined-rules into a mixed logit model, we further identify one key safe condition and four key risk conditions leading to serious crashes (i.e., fatalities and injuries). Each key risk condition is discussed and compared with its adjacent rules. Based on the findings, some countermeasures to rectify the freeway's serious one-vehicle crashes are proposed.

1. Introduction

A comprehensive understanding of the key risk conditions that may have contributed to different degrees of severity in vehicle crashes can facilitate the traffic engineers to initiate practical traffic safety programs. In the past, a number of works employed the parametric statistical methods to analyze crash severity—for example, binary outcome models (Al-Ghamdi, 2002; Sze and Wong, 2007; Helai et al., 2008; Lee and Abdel-Aty, 2008), ordered discrete outcome models (O'Donnell and Connor, 1996; Srinivasan, 2002; Tay and Rifaat, 2007; Rifaat and Chin, 2007; Pai and Saleh, 2007; Eluru et al., 2010; Zhu and Srinivasan, 2011) and unordered multinomial discrete outcome models (Shankar et al., 1996; McFadden and Train, 2000; Milton et al., 2008; Haleem and Abdel-Aty, 2010). Among them, ordered discrete outcome models have two main limitations including the constraint on the variable influence (e.g. a variable would either increase or decrease crash severity) and under-reporting, especially for low severity levels in accident data (Savolainen and Mannering, 2007; Yamamoto et al., 2008). The mixed logit model, a more generalized modeling approach, could account for heterogeneous effects and correlation in unobserved factors by allowing the random parameters to differ across crash-involved road users in a mixing distribution. Due to the limitation of designated distributions of parametric modeling, other distribution-free methods, such as decision tree (Chang and Chen, 2005; Chang and Wang, 2006) and artificial neural network (Chiou, 2006; Delen et al., 2006; Chimba and Sando, 2009), were also employed to analyze the crash severity. A comprehensive review of most recent accident models can be found in Savolainen et al. (2011). The aforementioned methods may be confronted by two difficulties. First, the parametric statistical methods may successfully identify the significant variables that can explain the crash severity or account for the complex relationships among them. The enumerated combination of significant variables, however, may be inadequate to explore the intertwined chained relationships, which can be crucial in analyzing vehicle crashes (Liu, 2007; Sze and Wong, 2007; Rhodes and Pivik, 2011). Generally, it is hard to presume the intertwined relationships among more than two variables; hence, the potential interactions among significant variables in crash severity study may not fully discover the crash causalities. Second, the classification outcomes resulted from decision tree or neural network methods are sometimes difficult to interpret. It can be ascribed to the hidden knowledge not fully explored from the crash dataset. In many circumstances, the prediction error of decision tree method is high, and the neural network method is functioning like a black box.

According to the error-chain theory, a typical vehicle crash can be resulted from a series of errors, not solely by a single factor. In this sense, mining the complicated rules to unveil the chain factors seems imperative and promising for crash severity studies.

* Corresponding author. Tel.: +886 2 23494940; fax: +886 2 23494953.
E-mail address: ycchiou@mail.nctu.edu.tw (Y.-C. Chiou).

0001-4575/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2012.05.017

Rule mining (a.k.a. rule generation, rule recovery, or classification/association rule mining) is one of the data mining techniques which search for useful knowledge from available database for better decision support. Rule mining is naturally modeled as multi-objective problems with three criteria: predictive accuracy, comprehensibility and interestingness (Freitas, 1999; Ghosh and Nath, 2004). Conventional rule mining models have intrinsic limitations on operational procedure or searching efficiency. In contrast, evolutionary rule mining algorithms provide more robust and efficient means to explore enormous search space. One such evolutionary algorithm is to use genetic algorithm (GA) to learn of the decision rules, termed genetic mining rule (GMR) (e.g. Freitas, 1999; Shin and Lee, 2002; Ghosh and Nath, 2004; Dehuri and Mall, 2006; Chen and Hsu, 2006). The performance of GMR algorithms has been proven and applied in many fields (Clarke et al., 1998; Chiou et al., 2010); yet the issue still arises as to conflicts and redundancies among the mined-rules.

This study aims to discover the key rules that potentially dominate the risk conditions causing crash severity, to accurately predict the crash severity, and moreover, to eradicate conflicts and redundancies among the mined-rules. The scope of the present study will limit the analysis of one-vehicle crashes in freeway contexts only. A two-stage mining framework is proposed to maximize the predictive accuracy of crash severity with a minimum number of key rules. In the first stage, a GMR model is proposed to identify the potential risk conditions that can best explain the degrees of crash severity. In the second stage, a mixed logit model is further estimated to test the significance of the mined risk conditions. The rest of this paper is organized as follows. Section 2 presents the crash data with definitions of contributing factors based on available dataset. Section 3 introduces the proposed mining framework with GMR model and mixed logit model. Section 4 presents the mining results by GMR model, which are further compared with those derived from the decision tree model. The estimation results of mixed logit model are also presented. Section 5 examines each of the key risk conditions and then proposes countermeasures accordingly. Finally, concluding remarks and suggestions for future research are addressed.

2. Data

The data were drawn from 2003 to 2007 National Traffic Accident Investigation Reports, provided by Taiwan National Police Agency. In the reports, each crash case has been carefully narrated by the police with digitized information about degrees of severity (fatal, injury, and property-damage only) of the involved parties, times of day of crash occurrence, vehicle movements (moving straight, right-turn, left-turn, lane-change), driver demographics (age, gender, driver sobriety), involved vehicle types, roadway geometrics, and other environmental conditions such as traffic control, weather (sunny, rain, fog, storm), pavement (wet, dry), lighting, among others. In view that more complicated and intertwined factors may exhibit in the collision cases involved with two or more parties, this paper only presents the one-vehicle crashes, which means that only a single vehicle is involved in the crash events. Two- or more than two-vehicle crashes will be studied in another paper.

During the five-year study period, a total of 5563 one-vehicle crash cases took place in Taiwan's freeways. Table 1 presents the detailed crash information, from which the potential 21 explanatory variables (recorded by the police) are defined. Each variable is categorical, with a brief description summarized in Table 1. Hereinafter, the most serious accidents are denoted as A1 (fatalities), followed by A2 (injuries), and A3 (property-damage only). The numbers of A1, A2 and A3 are 226, 1593, and 3744, respectively—a rather uneven distribution also seen in many other countries. To overcome the small number of observations in A1 crashes, which may lead to unreasonable mined-results by the GMR model, both A1 and A2 crashes are combined (1819 cases) and regarded as "serious crashes"; A3 crashes (3744 cases) are categorized as "minor crashes" in the following analysis. However, while estimating the mixed logit model, the three-level A1, A2 and A3 crashes are used to capture the statistical implications of risk conditions more precisely.

3. The mining framework

The core logic of the proposed mining framework contains two stages: (1) developing a GMR model to identify the key risk conditions and (2) formulating a mixed logit model to examine the significance of the risk conditions mined. In the first stage, the proposed GMR model is used to discover the "if-then" rules that can best elucidate the one-vehicle crash severity in the freeway contexts over the study horizon. For comparison, a decision tree (DT) model is also introduced to analyze the same dataset. In the second stage, the mixed logit model is formulated. Details of the two-stage modeling framework are depicted as follows.

3.1. The GMR model

The proposed GMR model contains encoding method, fitness function, genetic operators, and rule selection, narrated below.

3.1.1. Encoding method

To represent the relationship between explanatory variables and crash severity, a chromosome is used to represent each potential "if-then" rule. The associated conditions in the "if part" are antecedence part and those in the "then part" are consequent part. The antecedent part consists of at least 1 and at most 21 variables x_i selected from Table 1. The consequent part is composed by only one variable y, that is, severity degree. Due to the uneven distribution of three crash cases as explained, the severity variable y is redefined as serious crash (1: fatal or injury) and minor crash (2: property damage only).

Generally, a rule can be regarded as a knowledge representation of the form "If A then C," where A is a set of cases satisfying the conjunction of predicting attribute values and C is a set of cases with the same predicted severity degree. Specifically, a typical rule i can be expressed as Rule i: "If $x_1 = a_{i1}$ and $x_2 = a_{i2}$ . . .and $x_j = a_{ij}$ . . . and $x_{21} = a_{i21}$ then $y = g_i$," or in short, "If A_i then C_i," where a_{ij} is the categorical value of jth attribute variable and g_i is the value of classification variable in rule i. A_i and C_i are the sets of parties satisfying the antecedent part and consequent part of rule i, respectively.

By encoding a rule as a chromosome, each gene is used to represent a corresponding variable. In this study, there are 21 antecedent variables and one consequent variable, thus the length of a chromosome is 22. Each gene will take one of the categorical values of the corresponding variable. Because the ranges of all variables are different, the ranges of gene values will vary. In any circumstance, if a gene in a rule antecedent takes a value of 0, it represents the corresponding variable not considered by the rule.

3.1.2. Fitness function

The role of fitness function is to evaluate the quality of the rule numerically. An individual chromosome (a rule) with higher fitness function value has a higher probability being selected to reproduce the offspring. Shin and Lee (2002) adopted hit ratio (confidence), also known as predictive accuracy plus coverage (Kim and Han, 2003), as the fitness function. What should be emphasized here, however, is the performance of the entire rule set in lieu of the performance of each individual rule. Due to the potential conflict and redundancy among rules, a well-performed individual rule does not