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

Table 1
One-vehicle crash information in Taiwan's freeways (2003–2007).

Variable	Definition	Type	Description	Number of crashes		
				A1	A2	A3
X_1	Surface condition	Categorical	1, Dry;	167	1128	2403
			2, wet	59	465	1341
x_2	Signal control	Categorical	1, None;	222	1556	3644
			2, yes	4	37	100
x_3	Driver gender	Categorical	1, Male;	203	1326	3206
			2, female	23	267	538
x_4	Weather	Categorical	1, Sunny;	158	1061	2236
			2, cloudy;	16	105	277
			3, rain, storm, fog	52	427	1231
x_5	Obstacle	Categorical	1, None;	217	1513	3456
			2, work zone;	6	47	169
			3, others	3	33	119
x_6	Lighting condition	Categorical	1, daytime;	96	836	2150
			2, dawn or dusk;	5	52	105
			3, nighttime with illumination;	66	337	765
			4, nighttime without illumination	59	368	724
x_7	Speed limit	Categorical (discretized)	1, 110 km/h;	65	608	1296
			2, 100 km/h;	75	574	1444
			3, 90–70 km/h;	41	227	458
			4, 60–40 km/h	45	184	546
x_8	Road status	Categorical	1, Straight road;	201	1482	3456
			2, grade and curved road;	16	48	108
			3, tunnel, bridge, culvert, overpass;	4	35	106
			4, others	5	28	74
x_9	Marking	Categorical	1, Lane line with marker;	212	1531	3587
			2, lane line without marker;	4	15	53
			3, no lane-changing line;	5	35	56
			4, no lane line	5	12	48
x_{10}	Use of seat belt	Categorical	1, Seat belt fastened;	152	1539	3729
			2, seat belt not fastened;	27	22	12
			3, unknown	47	32	3
x_{11}	Use of cell phone	Categorical	1, Not in use;	127	1548	3725
			2, use;	0	4	7
			3, unknown	99	41	12
x_{12}	License	Categorical	1, With license;	202	1483	3608
			2, without license;	20	101	128
			3, unknown	4	9	8
x_{13}	Driver occupation	Categorical	1, In job;	123	1025	2640
			2, student;	24	59	108
			3, jobless;	15	93	119
			4, unknown	64	416	877
x_{14}	Driver age	Categorical (discretized)	1, Under 30 years old;	81	610	1313
			2, 30–40 years old;	62	426	1224
			3, 40–50 years old;	48	339	781
			4, 50–65 years old;	29	194	404
			5, above 65 years old	6	24	22
x_{15}	Time period	Categorical (discretized)	1, 07:01–09:00 morning peak;	18	111	304
			2, 09:01–16:00 daytime;	63	530	1357
			3, 16:01–19:00 afternoon peak;	19	205	532
			4, 19:01–23:00 nighttime;	34	235	528
			5, 23:01–07:00 midnight	92	512	1023
x_{16}	Location	Categorical	1, Traffic lane;	119	1153	2626
			2, shoulder;	55	226	527
			3, median;	9	17	29
			4, accelerating or decelerating lane, ramp;	33	160	425
			5, toll plaza and others	10	37	137
x_{17}	Vehicle type	Categorical	1, Passenger car;	145	994	2381
			2, light truck;	38	394	680
			3, bus;	3	17	40
			4, heavy truck, trailer and tractor;	36	166	617
			5, others (motorcycle and bicycle)	4	22	26

Table 1 (*Continued*)

Variable	Definition	Type	Description	Number of crashes		
				A1	A2	A3
x_{18}	Action	Categorical	1, Forward;	180	1272	3055
			2, left lane-change;	7	93	189
			3, right lane-change;	10	107	241
			4, abrupt deceleration;	3	40	117
			5, others	26	81	142
x_{19}	Alcoholic use	Categorical	1, No;	117	1262	3295
			2, under 0.25 mg/l (or 0.05%);	12	62	75
			3, over 0.25 mg/l (or 0.05%);	47	234	343
			4, cannot be tested;	37	15	20
			5, unknown	13	20	11
x_{20}	Journey purpose	Categorical	1, Commuting trip;	23	190	467
			2, business trip;	7	97	224
			3, transportation activity;	42	196	583
			4, visiting/shopping trip;	27	204	382
			5, others	127	906	2088
x_{21}	Major cause	Categorical	1, Improper lane-change;	2	77	172
			2, speeding;	49	168	496
			3, fail to keep a safe distance;	4	24	152
			4, alcoholic use;	41	226	323
			5, fail to pay attention to the front;	17	104	277
			6, other driver's liability;	87	699	1736
			7, factors not attributed to drivers	26	295	588
y	Severity	Categorical	1, Fatal;	226		
			2, injury;		1593	
			3, property-damage only			3744

necessarily imply that the combination of these rules will automatically perform well. Hence, this paper sets the fitness function as an increase of correctly classified cases by the rule set in ways that it combines the previously mined rules with the newly included rule, expressed as follows:

$$f_i^t = n_{S+R_i} - n_S \tag{1}$$

where f_i^t represents the incremental number of correctly classified cases if R_i is selected at the learning epoch t; n_S represents the number of cases correctly classified by the rule set S, which comprises the selected rules up to the learning epoch t; n_{S+R_i} represents the number of cases correctly classified by the rule set S and the rule i (R_i). In order to maximize the incremental increase in the number of correctly classified cases, the algorithm would avoid selecting the rules which are conflict or redundant to the previously selected rules in the rule set S. In so doing, the problem of conflict or redundancy would be effectively mitigated during the stepwise rule mining process. Additionally, because each selected rule is to maximally increase the correctly classified cases, the learning results based on the objective function Eq. (1) should be identical to those based in Eq. (2), which aims to maximize the predictive accuracy of all selected rules.

$$F_i^t = \frac{n_{S+R_i}}{N} \tag{2}$$

where F_i^t is the predictive accuracy rate at the epoch t if R_i is added to the incumbent rule; N represents the total number of crash cases.

Two performance indices – coverage and predictive accuracy – are also computed for rule comparison. The coverage of R_i is denoted by $CR_i = |A_i|$, representing the cardinality of set A_i (i.e., the set contains the crash cases satisfying the antecedent part of R_i). The predictive accuracy of R_i is denoted by $PA_i = |A_i \cap C_i|/|A_i|$ (Freitas, 1999), where C_i represents the set containing the cases satisfying the consequent part of R_i. $A_i \cap C_i$ represents the set containing the cases satisfying both antecedent and consequent parts of R_i; $|A_i \cap C_i|$ is the cardinality of the set $A_i \cap C_i$.

Since the genes in the proposed GMR model are not binary encoded, simple genetic algorithms proposed by Goldberg (1989) cannot be used. In turn, this study employs the following max-min-arithmetical crossover, proposed by Herrera et al. (1998) and the non-uniform mutation, proposed by Michalewicz (1992).

(1) Max-min-arithmetical crossover

Let $G_w^t = \{gw_1^t, \ldots, g_{wk}^t, \ldots, g_{wK}^t\}$ and $G_v^t = \{g_{v1}^t, \ldots, g_{vk}^t, \ldots, g_{vK}^t\}$ be two chromosomes selected for crossover, the following four offsprings can be generated:

$$G_1^{t+1} = aG_w^t + (1-a)G_v^t \tag{3}$$

$$G_2^{t+1} = aG_v^t + (1-a)G_w^t \tag{4}$$

$$G_3^{t+1} \text{ with } g_{3k}^{t+1} = \min\{g_{wk}^t, g_{vk}^t\} \tag{5}$$

$$G_4^{t+1} \text{ with } g_{4k}^{t+1} = \max\{g_{wk}^t, g_{vk}^t\} \tag{6}$$

where a is a parameter ($0 < a < 1$) and t is the number of generations.

(2) Non-uniform mutation

Let $G_t = \{g_1^t, \ldots, g_k^t, \ldots, g_K^t\}$ be a chromosome and the gene g_k^t be selected for mutation (the domain of g_k^t is $[g_k^l, g_k^u]$), the value of g_k^{t+1} after mutation can be computed as follows:

$$g_k^{t+1} = \begin{cases} g_k^t + \Delta(t, g_k^u - g_k^t) & \text{if } b = 0 \\ g_k^t - \Delta(t, g_k^t - g_k^l) & \text{if } b = 1 \end{cases} \tag{7}$$

where b randomly takes the binary value of 0 or 1. The function $\Delta(t, z)$ returns to a value in the range of $[0,z]$ such that the probability of $\Delta(t, z)$ approaches 0 as t increases:

$$\Delta(t, z) = z(1 - r^{(1-t/T)^h}) \tag{8}$$

where r is a random number in the interval $[0,1]$, T is the maximum number of generations and h is a given constant. In Eq. (8), the value returned by $\Delta(t, z)$ will gradually decrease as the evolution progresses.

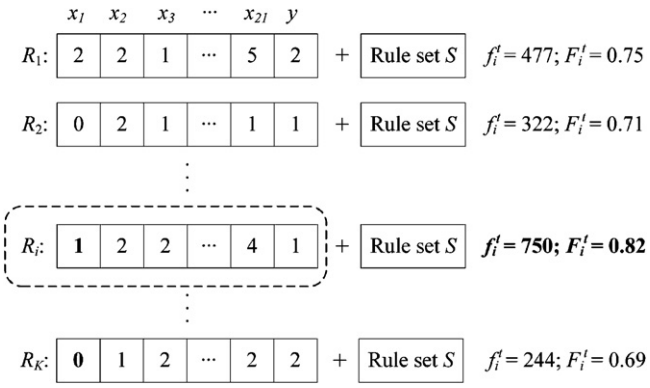

Fig. 1. The encoding and selection of rules.

3.1.3. Rule selection

Conventional GMR models simultaneously select a set of rules to achieve the objective function. It is inevitable that some mutually conflicting or redundant rules will be selected, which not only deteriorates the model performance but increases the difficulties in interpretation (in this study, interpreting the causal relationships between explanatory variables and crash severity). Moreover, the mined rules are often too complicated or too lengthy to be interpreted in a sensible way. To overcome these difficulties, instead of learning the rules simultaneously, this paper proposes a novel "stepwise rule mining approach," which contains the following steps:

Step 0: **Initialization**: Randomly generate the initial population and set $S = \Phi$ and $t = 1$.
Step 1: **Rule selection**: Select the rule (R_i) with maximum f_i^t as shown in Fig. 1.
Step 2: **Rule modification**: Perform rule modifications on R_i according to improvement and parsimony strategies and then add the modified rule to the incumbent rule set S. Let $t = t + 1$.
Step 2-1: **Improvement strategy**: Sequentially vary the value of each gene of R_i from x_1 to x_{21}. If the predictive accuracy is improved, then accept the modification. Otherwise, the rule remains unchanged.
Step 2-2: **Parsimony strategy**: Set the value of each gene of the candidate rule to be 0 in a sequence from x_1 to x_{21}, suggesting the corresponding factor not considered in the rule. If the predictive accuracy is not deteriorated, then accept the modification. Otherwise, the rule remains unchanged.
Step 3: **Stop condition**: Repeat steps 1 and 2 until the predictive accuracy of the incumbent rule set can no longer be improved by adding any other rules or the preset maximum number of epochs has reached.

If more than one rule with different consequent parts (i.e., different severity degrees) has been fired by the same crash case, the predicted severity degree should be determined by the consequent part of the rule with the highest predictive accuracy among all fired rules.

3.2. The mixed logit model

Recently, the mixed logit model, or random parameters logit model, has been applied for the analysis of crash injury-severity (e.g. Milton et al., 2008; Kim et al., 2010). The mixed logit addresses the limitations of the multinomial logit by allowing for heterogeneous effects and correlation in unobserved factors (Train, 2009). A mixed logit model is derived with the addition of a second error term to the severity function:

$$S_{in} = \beta_i X_{in} + [\eta_{in} + \varepsilon_{in}] \tag{9}$$

where S_{in} is the severity function determining the injury-severity level i (A1, A2, A3) on crash case n. β_i is a vector of parameters. X_{in} is a vector of explanatory variables. η_{in} is a random error term with zero mean. ε_{in} is the error term that is independent and identically distributed, and does not depend on underlying parameters or data.

The mixed logit allows the parameter vector β_i to vary across the crash-involved drivers. β_i may be either fixed or randomly distributed with fixed means, allowing for heterogeneous effects. A mixing distribution is introduced to the model formulation, resulting in injury severity probabilities as follows (Train, 2009):

$$P_{in}(i) = \int_X \frac{\exp[\beta_i X_{in}]}{\sum_i \exp[\beta_i X_{in}]} f(\beta|\phi) d\beta \tag{10}$$

where $f(\beta|\phi)$ is a density function of β and ϕ is a vector of parameters which describe the density function (mean and variance).

The functional form of the parameter density functions is given to normal distribution and a simulation-based maximum likelihood method with Halton draws can be applied in the model estimation (McFadden and Train, 2000).

In order to assess the effects of explanatory variables estimates on injury-severity outcome probabilities, elasticities are further computed. However, for categorical (dummy) explanatory variables, a regular elasticity cannot be calculated since the probability is not differentiable. To explore the marginal effect of such a categorical variable, a pseudo-elasticity, which gives the percent effect on the injury-severity probabilities of the variable switching from a value of 0 to 1, can be calculated as follows (Kim et al., 2010; Morgan and Mannering, 2011):

$$E_{x_{nk}}^{P_{in}} = \frac{P_{in}[\text{given } x_{nk} = 1] - P_{in}[\text{given } x_{nk} = 0]}{P_{in}[\text{given } x_{nk} = 0]} \tag{11}$$

where P_{in} is defined by Eq. (10), x_{nk} is the kth explanatory variable associated with injury severity i for crash case n. Direct and cross pseudo-elasticities are presented in this study as a measure of the marginal effect of an explanatory variable by taking the average over the whole sample.

4. Empirical results

In the following analysis, 70% of the crash cases are randomly chosen for training (i.e., 3895 cases) and the remaining 30% (1668 cases) are used for validation. A χ^2-test has shown that the severity distributions between training and validation do not differ significantly ($p < 0.05$).

4.1. The results of GMR model

The parameters of the proposed GMR model are set as follows: population size = 50, crossover rate = 0.85, mutation rate = 0.08. The learning process is depicted in Fig. 2.

Although the misclassification rate can be monotonically lowered by increasing the number of rules selected, a good GMR model should be able to fit the training data, but more importantly, to fit the validation data as well. According to Fig. 2, to avoid overtraining, 29 rules are thought appropriate because the validation misclassification rate of the rule set has reached the lowest value. Table 2 gives the prediction accuracies for both training and validation under various severity degrees. It shows that in the training dataset, the proposed GMR model can predict the serious crashes with a correct rate of 60.20% and minor crashes with a correct rate of 84.43%. The overall correct rates of the proposed GMR model in training and in validation are 76.51% and 74.82%, respectively.

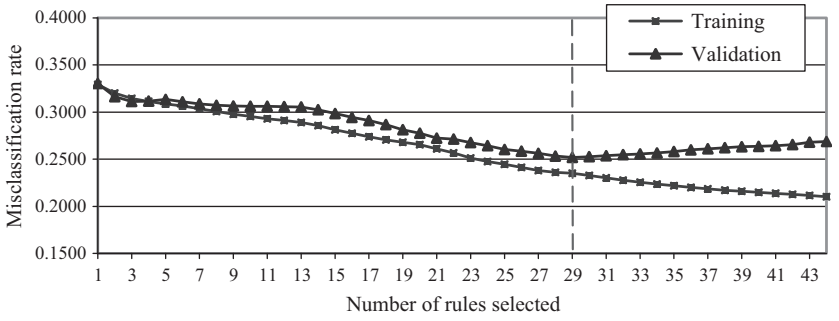

Fig. 2. Learning process of GMR model.

Table 2
Different severity cases (percentages) predicted by GMR model.

Datasets	Real severity	Predicted severity		Total
		Serious (A1 + A2)	Minor (A3)	
Training	Serious (A1 + A2)	**767 (60.20%)**	507 (39.80%)	1274 (100.00%)
	Minor (A3)	408 (15.57%)	**2213 (84.43%)**	2621 (100.00%)
	Total	1175	2720	3895
Validation	Serious (A1 + A2)	**325 (59.63%)**	220 (40.37%)	545 (100.00%)
	Minor (A3)	200 (17.81%)	**923 (82.19%)**	1123 (100.00%)
	Total	525	1143	1668

Note: The bold values represent the correctly predicted cases.

Table 3 reports the finally selected rules together with its corresponding performance indices (hereinafter, *CR* denotes coverage and *PA* denotes predictive accuracy). Note that the selected-rules have been ranked in a descending order according to predictive accuracy values.

Most of the rules in Table 3 can be readily inspected and explained by the "if-then" relationship of the rules themselves. Taking R_7 as an example, the rule indicates that "If the weather is sunny, speed limit is 110 km/h, and the car driver does not hold a valid license, then the crash tends to be serious (fatal or injury)." As for R_{11}, the rule says that "If the seat belt is not fastened, then the crash tends to be serious (fatal or injury)"; in contrast, R_{26} says that "If the seat belt is fastened, then the crash tends to be minor (property-damage only)."

It is interesting to note that the two performance indices, *CR* and *PA*, are negatively correlated (−0.66), suggesting that the higher coverage the rule has, the lower predictive accuracy it would be. Also note that the rules associated with minor crashes ($y = 2$) are all ranked behind the rules associated with serious crashes ($y = 1$). In general, the rules associated with minor crashes ($y = 2$) have

Table 3
Optimal combination of rules mined by GMR model.

Rules	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	y	CR_i	PA_i
R_1																	5		3			1	18	0.944
R_2			1											1	5						4	1	39	0.897
R_3				2									3									1	37	0.892
R_4								1								2			2			1	29	0.897
R_5			2													1		2				1	31	0.871
R_6	2				3		3															1	25	0.880
R_7				1			1					2					1					1	44	0.864
R_8							3							2			2					1	42	0.857
R_9													1		5	2	2					1	19	0.842
R_{10}	1		1												3				3			1	34	0.824
R_{11}										2												1	79	0.823
R_{12}						3							3	1	5							1	32	0.813
R_{13}								2									2		1			1	43	0.814
R_{14}		1					3	1											2			1	48	0.813
R_{15}																1	4		3			1	30	0.800
R_{16}							4							5	2							1	61	0.787
R_{17}			2										1		2		2					1	36	0.778
R_{18}				1												2					2	1	48	0.771
R_{19}						3			3							1						1	45	0.756
R_{20}						1						2					2					1	47	0.745
R_{21}							2							4							7	1	87	0.713
R_{22}						4								3			2					1	96	0.698
R_{23}						1	3						1									2	64	0.813
R_{24}				1				3														2	60	0.767
R_{25}											1										6	2	1015	0.697
R_{26}										1												2	3800	0.687
R_{27}																		1				2	3157	0.678
R_{28}					1																	2	3651	0.668
R_{29}				1					1			1										2	2199	0.653

Table 4
Different severity cases (percentages) predicted by DT model.

Datasets	Real severity	Predicted severity		Total
		Serious (A1 + A2)	Minor (A3)	
Training	Serious (A1 + A2)	**502 (39.40%)**	772 (60.60%)	1274 (100.00%)
	Minor (A3)	387 (14.77%)	**2234 (85.23%)**	2621 (100.00%)
	Total	889	3006	3895
Validation	Serious (A1 + A2)	**214 (39.27%)**	331 (60.73%)	545 (100.00%)
	Minor (A3)	177 (15.76%)	**946 (84.24%)**	1123 (100.00%)
	Total	391	1277	1668

Note: The bold values are the correctly predicted cases.

higher coverage and lower predictive accuracy than the rules associated with serious crashes ($y = 1$). The average coverage (predictive accuracy) of the rules for minor and serious crashes includes 1992 cases (70.90%) and 44 cases (82.17%), respectively. The coverage of the rules for serious crashes ranges from 18 to 96 cases, suggesting that the risk conditions of serious crashes are rather diverse. In other words, there will be no single risk condition that can explain noticeable percentages of the fatal or injury crashes in most circumstances.

4.2. The results of DT model

Via trial-and-error, the parameters of DT model are set as follows: splitting criterion is Gini reduction; minimum number of observations in a leaf is 1; observations required for a split search is 8; maximum number of branches from a node is 2; maximum depth of tree is 6; splitting rules saved in each node is 5. The model is executed by SAS Enterprise Miner Release 4.3 with learning process depicted in Fig. 3. Note that the misclassification rate decreases as the number of leaves gets larger.

A total of 18 rules have been generated by the DT model, 8 of which are associated with serious crashes ($y = 1$) and 10 associated with minor crashes ($y = 2$), summarized as follows:

R_1: If $x_{11} = 3$ Then $y = 1$.
R_2: If $x_{11} = 2$ Then $y = 2$.
R_3: If $x_{21} = 2$ and $x_{10} = \{2, 3\}$ and $x_{17} = \{1, 4\}$ and $x_{11} = 1$ Then $y = 1$.
R_4: If $x_3 = 2$ and $x_4 = \{2, 3\}$ and $x_{17} = \{2, 3, 5\}$ and $x_{11} = 1$ Then $y = 1$.
R_5: If $x_3 = 1$ and $x_4 = \{2, 3\}$ and $x_{17} = \{2, 3, 5\}$ and $x_{11} = 1$ Then $y = 2$.
R_6: If $x_{12} = 1$ and $x_{19} = 1$ and $x_{10} = 1$ and $x_{17} = \{1, 4\}$ and $x_{11} = 1$ Then $y = 2$.
R_7: If $x_{21} = \{2, 3, 4, 5, 7\}$ and $x_{19} = \{2, 3, 4, 5\}$ and $x_{10} = 1$ and $x_{17} = \{1, 4\}$ and $x_{11} = 1$ Then $y = 2$.
R_8: If $x_{15} = \{2, 4, 5\}$ and $x_{21} = \{1, 3, 4, 5, 6, 7\}$ and $x_{10} = \{2, 3\}$ and $x_{17} = \{1, 4\}$ and $x_{11} = 1$ Then $y = 1$.
R_9: If $x_{15} = \{1, 3\}$ and $x_{21} = \{1, 3, 4, 5, 6, 7\}$ and $x_{10} = \{2, 3\}$ and $x_{17} = \{1, 4\}$ and $x_{11} = 1$ Then $y = 2$.
R_{10}: If $x_{13} = \{1, 2, 4\}$ and $x_{21} = \{2, 3, 6\}$ and $x_4 = 1$ and $x_{17} = \{2, 3, 5\}$ and $x_{11} = 1$ Then $y = 2$.

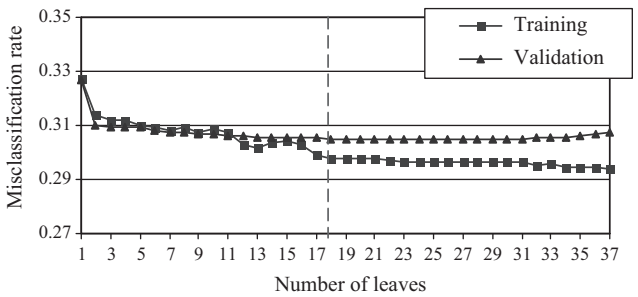

Fig. 3. Learning process of DT model.

R_{11}: If $x_{13} = 3$ and $x_{21} = \{2, 3, 6\}$ and $x_4 = 1$ and $x_{17} = \{2, 3, 5\}$ and $x_{11} = 1$ Then $y = 1$.
R_{12}: If $x_{20} = 3$ and $x_{21} = \{1, 4, 5, 7\}$ and $x_4 = 1$ and $x_{17} = \{2, 3, 5\}$ and $x_{11} = 1$ Then $y = 2$.
R_{13}: If $x_{21} = \{1, 2, 3, 6, 7\}$ and $x_{12} = \{2, 3\}$ and $x_{19} = 1$ and $x_{10} = 1$ and $x_{17} = \{1, 4\}$ and $x_{11} = 1$ Then $y = 2$.
R_{14}: If $x_{21} = 5$ and $x_{12} = \{2, 3\}$ and $x_{19} = 1$ and $x_{10} = 1$ and $x_{17} = \{1, 4\}$ and $x_{11} = 1$ Then $y = 1$.
R_{15}: If $x_{14} = \{1, 2\}$ and $x_{21} = \{1, 6\}$ and $x_{19} = \{2, 3, 4, 5\}$ and $x_{10} = 1$ and $x_{17} = \{1, 4\}$ and $x_{11} = 1$ Then $y = 1$.
R_{16}: If $x_{14} = \{2, 3, 5\}$ and $x_{21} = \{1, 6\}$ and $x_{19} = \{2, 3, 4, 5\}$ and $x_{10} = 1$ and $x_{17} = \{1, 4\}$ and $x_{11} = 1$ Then $y = 2$.
R_{17}: If $x_{15} = \{1, 2, 3, 4\}$ and $x_{20} = \{1, 2, 4, 5\}$ and $x_{21} = \{1, 4, 5, 7\}$ and $x_4 = 1$ and $x_{17} = \{2, 3, 5\}$ and $x_{11} = 1$ Then $y = 2$.
R_{18}: If $x_{15} = 5$ and $x_{20} = \{1, 2, 4, 5\}$ and $x_{21} = \{1, 4, 5, 7\}$ and $x_4 = 1$ and $x_{17} = \{2, 3, 5\}$ and $x_{11} = 1$ Then $y = 1$.

Table 4 further presents the number of cases with different severity degrees predicted by the DT model. In predicting the minor crashes, the DT model performs slightly better (the correct rates in training and validation are 85.35% and 84.24%, respectively, as indicated in Table 4) than the proposed GMR model (84.43% and 82.19%, respectively, as indicated in Table 3). However, in predicting the serious crashes, the DT model performs far inferior (the correct rates in training and validation are 39.40% and 39.27% respectively as shown in Table 4) to the proposed GMR model (60.20% and 59.63% respectively as indicated in Table 3). The overall correct rates of the GMR model outperform in both training and validation (76.51% and 74.82%, respectively) as opposed to 70.24% and 69.54% respectively for the DT model. As such, the subsequent mixed logit model in the second stage will be estimated from the learning results of the GMR model, in lieu of the DT model.

4.3. The mixed logit model

The 29 rules (risk conditions), mined by the GMR model, are set as the explanatory (dummy) variables to explain the crash severity. The estimation results are presented in Table 5 with correct signs. The parameters found to be random were the constant and R_{11} contributing to fatal crash, implying that "seat belt not fastened" (risk condition R_{11}) tends to be risky but the effects can vary across the drivers. The constant for the fatal proportion is normally distributed with mean 1.694 and standard deviation (σ_c) 1.368. It suggests that, in probability, 10.8% of the fatalities have constant term less than 0 and 89.2% greater than 0. Similarly, the risk condition on "seat belt not fastened" R_{11} for the fatal proportion is normally distributed with mean 1.941 and standard deviation ($\sigma_{R_{11}}$) 1.032. Therefore, in probability, 3% of the fatalities have R_{11} less than 0 and 97% greater than 0. Moreover, the coefficients of R_{25} and R_{27} are not significant in fatal and injury crash levels, as opposed to property-damage only ($p < 0.05$).

Table 5 also presents the detailed pseudo-elasticity values, including own-elasticity and cross-elasticity effects, associated

Table 5
Estimation results of mixed logit model using the 29 rules mined by GMR model.

Variable	Fatal					Injury				
	Coefficient	*t*-Statistic	Elasticity			Coefficient	*t*-Statistic	Elasticity		
			Fatal	Injury	Property damage only			Fatal	Injury	Property damage only
Constant	1.694	4.569*				0.887	2.143			
σ_c	1.368	2.050								
***R*$_1$**	3.390	4.977*	**83.0%**	−3.3%	−4.3%					
***R*$_2$**						2.653	3.416*	−3.6%	**53.1%**	−4.9%
R_3						3.342	3.169*	−1.8%	48.3%	−2.0%
***R*$_4$**						2.486	3.825*	−4.3%	**91.5%**	−6.5%
R_5						3.269	3.096*	−2.6%	27.4%	−2.8%
***R*$_6$**	3.915	3.186*	**69.4%**	−1.1%	−2.1%	3.138	2.830*	−6.2%	35.2%	−5.5%
R_7						2.094	2.611*	−0.6%	10.9%	−0.9%
R_8	3.440	3.340*	45.7%	−2.1%	−3.1%	2.641	3.328*	−3.3%	21.2%	−3.7%
R_9						2.126	2.342	−2.2%	14.1%	−2.3%
R_{10}						2.121	2.692*	−2.4%	9.1%	−2.6%
R_{11}	1.941	2.229	8.3%	0.5%	0.6%					
σ_{R11}	1.032	2.017								
R_{12}						1.877	2.708*	−0.3%	5.1%	0.1%
R_{13}						1.972	3.357*	−0.4%	7.2%	−0.7%
R_{14}	2.115	2.122	11.5%	−1.8%	−1.7%	2.110	3.182*	−2.3%	14.2%	−4.7%
R_{15}						1.519	2.081	−0.2%	3.1%	−0.4%
R_{16}						1.827	3.039*	−1.2%	6.3%	−1.6%
R_{17}						2.027	3.011*	−1.4%	18.2%	−1.5%
R_{18}	2.654	3.301*	13.7%	−2.2%	−4.2%	1.634	2.560	−0.2%	4.2%	−0.4%
R_{19}						2.132	3.447*	−3.5%	17.3%	−4.7%
R_{20}						1.897	3.415*	−0.3%	7.2%	−0.5%
R_{21}						1.511	3.784*	−0.5%	5.4%	−0.8%
R_{22}						1.705	4.535*	−0.5%	12.5%	−0.9%
R_{23}	−2.061	−1.875	−45.3%	3.1%	1.1%	−0.888	−2.428	3.2%	−16.2%	3.3%
R_{24}						−0.755	−2.061	1.1%	−8.1%	1.3%
***R*$_{26}$**	−4.916	−12.814*	**−510.3%**	35.7%	15.4%	−2.416	−6.388*	39.6%	**−175.0%**	70.5%
R_{28}						−0.3917	−2.321	−6.5%	−27.6%	−11.9%
R_{29}						−0.2158	−2.798*	−2.2%	−19.0%	−4.1%
Log-likelihood	−2646.808									
Restricted log-likelihood	−4279.095									
Likelihood ratio index	0.381									

Note: 1: The coefficients of R_{25} and R_{27} are not significant ($p < 0.05$).
2: The bold variables have pseudo-elasticity values greater than 50% or less than −100%.
3: σ denotes the standard deviation of distribution.
* Indicates the parameter statistically significant at the 0.01 level; whereas unmarked parameters indicate statistically significant at the 0.05 level.

with the 29 risk conditions mined, which give the percent effects on the injury-severity probabilities of the variables switching from 0 to 1. In theory, if an elasticity value of a risk condition is greater (less) than 100%, then the corresponding risk condition should be regarded as sensitive (insensitive) to fatality or injury. However, to explore the marginal effects of categorical variables in a stricter manner, we regard the pseudo-elasticity values less than 50% as "little sensitive," 50–100% as "likely sensitive," and greater than 100% as "sensitive" to fatal and/or injury. As such, five risk conditions in Table 5 have been identified with own-elasticity values greater than 50%, suggesting that these risk conditions have potentially sensitive effects on serious crashes (fatalities and/or injuries). Specifically, R_1 has a pseudo-elasticity value of 83.0% on fatal, R_2 has a value of 53.1% on injury, R_4 has a value of 91.5% on injury, R_6 has a value of 69.4% on fatal, and R_{26} has a value of −510.3% on fatal and a value of −175.0% on injury. However, the negative signs of R_{26} suggest that "use of seat belt" is the most critical variable in reducing serious crashes; therefore, it should be regarded as a "key safe condition," as opposed to the remaining four "key risk conditions" R_1, R_2, R_4, and R_6.

For comparison, another mixed logit model simply using the original 21 explanatory variables is also attempted. Since the original explanatory variables are all categorical, various numbers of dummy variables must be introduced (the number of dummy variables = the number of categories − 1), making 62 dummy explanatory variables for the initial estimation. By excluding insignificant variables, the estimation results of the mixed logit model with the original 21 variables are reported in Table 6. Most of the signs are also reasonable but the parameters found to be random were "cell phone not in use" contributing to injury crash, implying that "cell phone not in use" tends to lower the crash severity and the effects vary across the drivers. It is also normally distributed with mean −2.141 and standard deviation 2.332, which results in 82.1% of the distribution less than 0, and 17.9% of the distribution greater than 0. Moreover, the variable "use of seat belt" in Table 6 has a pseudo-elasticity value of −336.5% on fatal and a value of −121.9% on injury, while the variable "use of cell phone" has a value of −382.3% on fatal. It indicates these two original explanatory variables are the most critical factors leading to serious crashes.

5. Discussions

We compare the estimation results presented in Tables 5 and 6. The log-likelihood ratio test shows that the model using the original 21 explanatory variables is significantly inferior to the model using the 29 mined-rule explanatory variables ($p < 0.05$). Consequently, the following discussion will be based on the GMR mined-rule estimation results to identify the key risk conditions.

According to Table 5, five key conditions R_1, R_2, R_4, R_6 and R_{26} have "likely sensitive" or "sensitive" effects on the crash severity. Therefore, these "key risk conditions" can be regarded

Table 6
Estimation results of mixed logit model using the original 21 explanatory variables.

Variable	Fatal					Injury				
	Coefficient	*t*-Statistic	Elasticity			Coefficient	*t*-Statistic	Elasticity		
			Fatal	Injury	Property damage only			Fatal	Injury	Property damage only
Constant	3.946	7.074*				3.815	5.773*			
Speed limit (110 km/h)						0.432	2.753*	−2.8%	6.1%	−2.5%
Speed limit (90–70 km/h)						0.569	2.539	−1.4%	3.1%	−1.4%
Marking (lane line without marker)						1.026	2.167	−0.3%	0.6%	−0.4%
Use of seat belt (fastened)	−3.570	−7.593*	**−336.5%**	7.2%	11.8%	−2.972	−5.045*	47.5%	**−121.9%**	32.5%
Use of cell phone (not in use)	−4.064	−11.192*	**−382.3%**	5.8%	12.0%	−2.141	−4.500*	20.2%	−37.7%	18.2%
σ Use of cell phone (not in use)						2.332	2.920*			
Driver occupation (jobless)						1.091	3.014*	−1.1%	1.5%	−0.9%
Driver age (under 30 years old)						0.418	2.364	−2.6%	6.2%	−2.3%
Driver age (40–50 years old)						0.384	1.997	−1.4%	3.4%	−1.2%
Driver age (50–65 years old)						0.575	2.290	−1.2%	2.6%	−0.9%
Driver age (above 65 years old)	1.661	2.767*	1.3%	−0.2%	−0.2%					
Location (shoulder, edge)	0.929	4.051*	12.4%	−0.7%	−1.1%					
Location (median)	1.841	2.953*	1.4%	−0.2%	−0.3%					
Location (accelerating or decelerating lane, ramp)						−1.114	−2.437	0.5%	−1.9%	0.4%
Vehicle type (passenger car)						−0.634	−3.080*	6.6%	−17.0%	5.6%
Vehicle type (heavy truck, trailer truck, tractor)						−1.399	−3.688*	2.4%	−10.0%	2.3%
Action (right lane-change)	−1.112	−2.180	−7.0%	0.2%	0.2%					
Alcoholic use (cannot be tested)	2.784	6.378*	1.8%	−1.3%	−1.5%					
Major cause (speeding)						−0.854	−3.269*	1.6%	−5.2%	1.4%
Major cause (fail to keep a safe distance)						−1.892	−3.167*	0.5%	−3.3%	0.5%
Major cause (other driver's liability)						−0.395	−2.619*	2.9%	−7.9%	2.6%
Log-likelihood	−2685.295									
Restricted log-likelihood	−4279.095									
Likelihood ratio index	0.372									

Note: 1: σ denotes the standard deviation of distribution.
2: The bold variables have pseudo-elasticity values greater than 50% or less than −100%.
* Indicates the parameter is statistically significant at the 0.01 level and unmarked parameters are statistically significant at the 0.05 level.

as the antecedent part of R_1, R_2, R_4, and R_6. In contrast, the "key safe conditions" can be regarded as the antecedent part of R_{26}. The "key safe condition" reflected by R_{26} obviously depicts that "fastening seat belt can reduce the crash severity once a driver gets involved in a crash." We further look into the remaining four "key risk conditions": R_1, R_2, R_4, and R_6. Hereinafter, the terms "Ratio" as shown in Table 7 is defined as the number of serious crashes divided by the number of total crashes. For comparison, the ratios of other rules adjacent to the four key risk conditions are also presented. The so-called "adjacent rules" are those rules using the same explanatory variables as the key rules in the antecedent parts, but one of the explanatory variables takes a different value. To save space, only the adjacent rules with ratio values greater than the ratio of each key rule will be reported. For instance, we have attempted 9 adjacent rules for R_1, but only 4 adjacent rules with ratios greater than 32.70% = (226 + 1593)/(226 + 1593 + 3744) are reported and denoted as $R_1 - 1 \sim R_1 - 4$.

R_1 contains two variables: "vehicle type = others" and "alcoholic use = over 0.25 mg/l." The vehicle type belonging to "others" category includes motorcycles and bicycles, which are not allowed to enter the freeways according to the regulations, because they are much more vulnerable than larger vehicles. Once the motorcyclists or bicyclists have alcoholic use, they might unconsciously enter the freeways and probably causing one-vehicle crash with serious severity, although such cases have rarely happened. Table 7 compares the ratios of the number of serious crashes to the number of total crashes of R_1 with its adjacent rules. Note that a total of 18 such cases are covered by this rule and almost all such crashes are serious (94.44%), much higher than its adjacent rules $R_1 - 1 \sim R_1 - 4$. No matter which types the vehicle may be, the crashes tend to be serious once the cyclists have alcoholic use. Accordingly, more intensive patrol is suggested to eradicate the alcohol-used cyclists illegally entering the freeways. Meanwhile, the geometrics and signs near the on-ramp areas should be designed in such a way to prevent motorcyclists or bicyclists from illegally entering the freeways.

R_2 contains four variables: "Driver gender = male," "Driver age = under 30 years old," "Time period = midnight" and "Major cause = alcoholic use." Table 8 compares the ratios of R_2 with its adjacent rules. Note that the selected rule has much higher ratio than its adjacent rules. Also note that the crashes caused by drunk young drivers in midnight or night time tend to have a higher

Table 7
Comparison of R_1 with its adjacent rules.

Rule	Antecedent part		Number of crashes			
	Vehicle type	Alcoholic use	Serious	Minor	Total	Ratio
R_1	Others	Over 0.25 mg/l	17	1	18	94.44%
$R_1 - 1$	Heavy truck, trailer and tractor	Over 0.25 mg/l	14	9	23	60.87%
$R_1 - 2$	Bus	Over 0.25 mg/l	1	1	2	50.00%
$R_1 - 3$	Light truck	Over 0.25 mg/l	27	30	57	47.37%
$R_1 - 4$	Passenger car	Over 0.25 mg/l	225	303	528	42.61%

Note: Ratio = the number of serious crashes/the number of total crashes, same for the remaining tables.

Table 8
Comparison of R_2 with its adjacent rules.

Rule	Antecedent part			Number of crashes			
	Driver age	Time period	Major cause	Serious	Minor	Total	Ratio
R_2	<30	Midnight	Alcoholic use	35	4	39	89.74%
R_2-1	30–40	Midnight	Alcoholic use	7	5	12	58.33%
R_2-2	40–50	Midnight	Alcoholic use	4	6	10	40.00%
R_2-3	<30	Night time	Alcoholic use	2	3	5	40.00%
R_2-4	<30	Midnight	Fail to pay attention to the front	4	6	10	40.00%
R_2-5	<30	Midnight	Factors not attributed to drivers	4	6	10	40.00%

Note: Driver gender was examined and no significant difference has been found.

Table 9
Comparison of R_4 with its adjacent rules.

Rule	Antecedent part			Number of crashes			
	Road status	Location	Alcoholic use	Serious	Minor	Total	Ratio
R_4	Straight road	Shoulder	Under 0.25 mg/l	26	3	29	89.66%
R_4-1	Straight road	Shoulder	Over 0.25 mg/l	14	4	18	77.78%
R_4-2	Straight road	Toll plaza and others	Under 0.25 mg/l	2	2	4	50.00%
R_4-3	Grade and curved road	Shoulder	Under 0.25 mg/l	1	1	2	50.00%
R_4-4	Straight road	Traffic lane	Under 0.25 mg/l	20	35	55	36.36%

Table 10
Comparison of R_6 with its adjacent rules.

Rule	Antecedent part			Number of crashes			
	Surface condition	Obstacle	Speed limit	Serious	Minor	Total	Ratio
R_6	Wet	Others	90–70 km/h	22	3	25	88.00%
R_6-1	Wet	No	90–70 km/h	79	143	222	35.59%

severity. Accordingly, in addition to more intense drunk driving enforcement, improving the illumination is perhaps an effective countermeasure to mitigate the risk of driving at night.

R_4 contains three variables: "Road status = straight road," "Location = shoulder or edge" and "Alcoholic use = under 0.25 mg/l." Table 9 compares the ratios of R_4 with its adjacent rules. Note that the selected rule still exhibits the highest ratio of serious crashes, followed by R_4-1 which is the same conditions with the selected-rule except that the drivers have even heavier alcoholic use. From Table 9, most of rules with higher severe crash ratio are involved with alcoholic use, suggesting the importance of enforcement in eradicating the drink-drive phenomena.

Similar to R_2, the influence of alcoholic use is also involved in R_4. The implications of R_4 mainly focus on the influence from environmental or spatial characteristics with alcoholic use, while those of R_2 mainly from driver or temporal characteristics with alcohol use. Synthesizing the analysis of R_2 and R_4, both are important references as to proposing the countermeasures for reducing the crash severity involved with drink-drive.

R_6 contains three variables: "Surface condition = wet," "Obstacle = others (e.g. fallen object, breakdown vehicles)" and "Speed limit = 90–70 km/h." The speed limit is 110 or 100 km/h for most segments in Taiwan's freeways, a reduced speed limit = 90–70 km/h indicates that the segments can be curved or in tunnels, toll stations or work zones. Table 10 compares the ratios of R_6 with its adjacent rules. The selected-rule suggests that the conditions with wet surface and unexpected obstacles in the reduced speed limit segments could be riskier than other segments. If there have no obstacles, the ratio can then be largely reduced to 35.59% as indicated by R_6-1. Even so, the wet surface at the reduced speed limit segments is still potentially dangerous. Accordingly, prompt removals of roadway fallen objects or breakdown vehicles and provisions of real-time incident information should be regarded as the possible countermeasures.

In addition to the abovementioned key risk conditions, other risk conditions are also briefly discussed as follows. R_{10} contains "heavy alcoholic use" and "afternoon peak." Once again, it suggests the necessity of law enforcement on drink-drive. R_{14} reveals a higher risk of light alcoholic use and cell phones usage under 90–70 km/h speed limit. This also supports the policy of forbidding cell phones usage during driving, especially in the reduced speed limit segments. As for R_{16}, it reveals that the elder drivers are in higher risk when they pass through the reduced speed limit segments. Thus, more visible speed limit signs or other safety devices are essential, especially for the aged drivers. R_{20} is composed of "without license" and "truck." It is understandable that truck drivers without licenses can jeopardize themselves and others. This can be eradicated through intensive law enforcement with higher fines. As for R_{22}, it reveals that truck drivers aged 40–50 may have higher risk during night-time without illumination.

It should be pointed out that few mined-rules in this study are rather difficult to interpret. For instance, R_{19}, crashes in night-time with illumination and without lane-changing behaviors tend to be serious. There have no manifest risk conditions to be identified here; nonetheless, it could be due to other factors (e.g. fatigue driving) not narrated by the police in the traffic accident investigation reports. This may call for further studies.

It is worth noting that crash data usually suffer from under-reporting effects, especially for the lower injury severity cases. An outcome-based sample, which is over-represented by accidents of higher severity, would result in biased parameters which skew the inferences on the effects of key safety variables (Yamamoto et al., 2008). Thus, more careful investigation into accident reports should be taken prior to models estimation.

6. Conclusions

This paper has contributed to propose a two-stage mining framework with the genetic mining rule (GMR) model and the mixed logit model to identify the joint effects of key risk conditions contributing to one-vehicle crash severity in freeway contexts. A novel stepwise rule mining algorithm is proposed to avoid selecting

conflicting or redundant rules. The empirical analysis based on the 2003–2007 one-vehicle crash cases occurred in Taiwan's freeways (A1: 226 cases, A2: 1593 cases, A3: 3744 cases has mined 29 rules in the first stage, which can achieve overall correct rates of 76.51% in training and 74.82% in validation). By incorporating these 29 mined-rules into a mixed logit model as explanatory dummy variables in the second stage, five key risk conditions leading to serious crashes have been identified. In-depth investigations on these key risk conditions are further discussed and some countermeasures to ameliorate the traffic safety are proposed accordingly. The empirical results have demonstrated that the proposed two-stage mining framework can satisfactorily identify the key risk conditions on crash severity.

Some directions for future studies can be identified. First, other information such as driver's mentality conditions, reaction behaviors, fatigues, and traffic conditions while crashes occurred are not recorded in the traffic accident investigation reports; but such information may be important for crash severity analysis as well. Thus, how to refine the police's traffic accident investigation reports, not only for liability purposes but also for research and practical traffic management purposes, can be an important topic for future study. Second, to capture the influence of each original explanatory variable on different risk conditions, it can be useful to establish a comprehensive relationship between individual explanatory variables and crash severity. Third, in most countries, fatal crash takes only a small portion of all crashes, but its contributing factors are of major concern. Our proposed GMR model used "coverage" as one of the performance indices, making it difficult in selecting the rules representing the most severe crashes due to the small sample size of fatal crash cases. We thus grouped the crashes into "serious" (fatal and injury) and "minor" (property damage only) and then mined the potential risk conditions contributing to the serious crashes. To enrich the fatal crash cases by studying a longer period of time will enhance the proposed GMR model. Last but not least, analysis of two-vehicle or more-than-two-vehicle crashes can be more challenging calling for another study, which requires developing a more sophisticated mining framework for use.

Acknowledgements

The authors are deeply indebted to two anonymous reviewers for their insightful comments and very constructive suggestions, which help correct several weak points in the original manuscript. This study was financially sponsored by the ROC National Science Council (NSC 97-2628-E-009-035-MY3).

References

Al-Ghamdi, A., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis and Prevention 34 (6), 729–741.
Chang, L.Y., Chen, W.C., 2005. Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research 36 (4), 365–375.
Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38 (5), 1019–1027.
Chen, T.C., Hsu, T.C., 2006. A GAs-based approach for mining breast cancer pattern. Expert Systems with Applications 30 (4), 674–681.
Chimba, D., Sando, T., 2009. Neuromorphic prediction of highway injury severity. Advances in Transportation Studies 19 (1), 17–26.
Chiou, Y.C., 2006. An artificial neural network-based expert system for the accident appraisal of two-car crash accidents. Accident Analysis and Prevention 38 (4), 777–785.
Chiou, Y.C., Lan, L.W., Chen, W.P., 2010. Contributory factors to crash severity in Taiwan freeways: genetic mining rule approach. Journal of the Eastern Asia Society for Transportation Studies 8, 1865–1877.
Clarke, D.D., Forsyth, R.S., Wright, R.L., 1998. Behavioural factors in accidents at road junctions: the use of a genetic algorithm to extract descriptive rules from police case files. Accident Analysis and Prevention 30 (2), 223–234.
Dehuri, S., Mall, R., 2006. Prediction and comprehensible rule discovery using a multi-objective genetic algorithm. Knowledge-Based System 19 (6), 413–421.
Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using series of artificial neural networks. Accident Analysis and Prevention 38 (3), 434–444.
Eluru, N., Paleti, R., Pendyala, R.M., Bhat, C.R., 2010. Modeling multiple vehicle occupant injury severity: a Copula-based multivariate approach. Transportation Research Record 2165, 1–11.
Freitas, A.A., 1999. On rule interestingness measures. Knowledge-Based Systems 12 (5–6), 309–315.
Ghosh, A., Nath, B., 2004. Multi-objective rule mining using genetic algorithms. Information Sciences 163 (1–3), 123–133.
Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, New York.
Haleem, K., Abdel-Aty, M., 2010. Examining traffic crash injury severity at unsignalized intersections. Journal of Safety Research 41 (4), 347–357.
Helai, H., Chor, C., Haque, M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 40 (1), 45–54.
Herrera, F., Lozano, M., Verdegay, J.L., 1998. A learning process for fuzzy control rules using genetic algorithms. Fuzzy Sets and Systems 100, 143–158.
Kim, M.J., Han, I., 2003. The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. Expert Systems with Applications 25 (4), 637–646.
Kim, J., Ulfarsson, G., Shankar, V., Mannering, F., 2010. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. Accident Analysis and Prevention 42 (6), 1751–1758.
Lee, C., Abdel-Aty, M., 2008. Presence of passengers: does it increase or reduce driver's crash potential? Accident Analysis and Prevention 40 (5), 1703–1712.
Liu, B.-S., 2007. Association of intersection approach speed with driver characteristics, vehicle type and traffic conditions comparing urban and suburban areas. Accident Analysis and Prevention 39 (3), 216–223.
McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. Journal of Applied Econometrics 15 (5), 447–470.
Michalewicz, Z., 1992. Genetic Algorithms + Data Structures = Evolution Programs. Springer, Berlin.
Milton, J., Shankar, V., Mannering, F., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention 40 (1), 260–266.
Morgan, A., Mannering, F., 2011. The effects of road-surface conditions, age, and gender on driver-injury severities. Accident Analysis and Prevention 43 (5), 1852–1863.
O'Donnell, C.J., Connor, D.H., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accident Analysis and Prevention 28 (6), 739–753.
Pai, C.W., Saleh, W., 2007. Exploring motorcyclist injury severity resulting from various crash configurations at T-junctions in the UK–an application of the ordered probit models. Traffic Injury Prevention 8 (1), 62–68.
Rifaat, S.M., Chin, H.C., 2007. Factors contributing to the severity of intersection crashes. Journal of Advanced Transportation 41 (1), 91–114.
Rhodes, N., Pivik, K., 2011. Age and gender differences in risky driving: the roles of positive affect and risk perception. Accident Analysis and Prevention 43 (3), 923–931.
Savolainen, P.T., Mannering, F.L., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accident Analysis and Prevention 39 (6), 955–963.
Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention 43 (5), 1666–1676.
Shankar, V., Mannering, F., Barfield, W., 1996. Effect of roadway geometrics and environmental factors on rural accident frequencies. Accident Analysis and Prevention 28 (3), 371–389.
Shin, K.S., Lee, Y.J., 2002. A genetic algorithm application in bankruptcy prediction modeling. Expert Systems with Applications 23 (3), 321–328.
Srinivasan, K., 2002. Injury severity analysis with variable and correlated thresholds: ordered mixed logit formulation. Transportation Research Record 1784, 132–142.
Sze, N.N., Wong, S.C., 2007. Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accident Analysis and Prevention 39 (6), 1267–1278.
Tay, R., Rifaat, S.M., 2007. Factors contributing to the severity of intersection crashes. Journal of Advanced Transportation 41 (3), 245–265.
Train, K., 2009. Discrete Choice Methods with Simulation. Cambridge University Press, NewYork, NY.
Yamamoto, T., Hashiji, J., Shankar, V., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accident Analysis and Prevention 40 (4), 1320–1329.
Zhu, X., Srinivasan, S., 2011. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Accident Analysis and Prevention 43 (1), 49–57.