Contributory Factors to Crash Severity in Taiwan's Freeways: Genetic Mining Rule Approach

Yu-Chiun CHIOU
Associate Professor
Institute of Traffic and Transportation
National Chiao Tung University
4F, 118 Sec.1, Chung Hsiao W. Rd., Taipei, Taiwan 10012
Fax: +886-2- 2349-4940
E-mail: ycchiou@mail.nctu.edu.tw

Lawrence W. LAN
Professor, Department of Global Marketing and Logistics; Dean, College of Management, MingDao University
Emeritus Professor, National Chiao Tung University
369 Wen-Hua Rd., Peetow, Changhua, Taiwan 52345
Fax: +886-4-887-9013
E-mail: lawrencelan@mdu.edu.tw

Wen-Pin CHEN
Ph.D. Candidate
Institute of Traffic and Transportation,
National Chiao Tung University,
4F, 118 Sec.1, Chung Hsiao W. Rd., Taipei, Taiwan 10012
Fax: +886-2-2909-7032
E-mail: wpchen.tt94g@nctu.edu.tw

Abstract: A crash is often caused by a series of errors and also attributed to a number of categorical explanatory factors. To explore the key rules that determine the most contributing factors to crash severity, this paper develops a novel genetic mining rule (GMR) model, which accounts for the conflict and redundancy of rules mined. To avoid over-mining caused by unevenly distributed data across different types of accidents, identical numbers of A1-type (fatal), A2-type (injury), and A3-type (non-injury) crash cases drawn from 2003-2007 Taiwan's freeway accident investigation reports are used for the analysis. A total of 39 rules are mined which can achieve an overall correct rate of 74.25% in training and 70.79% in validation, respectively, much higher than those yielded by the decision tree model. Travel period, major cause, collision type and journey purpose are found as the four major contributory factors to crash severity in this study.

Key Words: *Crash data analysis, crash severity, decision tree, genetic mining rule.*

1. INTRODUCTION

Crash data analysis can be carried out by two main approaches: collective approach and individual approach (Abdel-Aty and Pande, 2007). The collective approach is characterized by crash frequency modeling. Frequency of crashes is aggregated over specific time periods (months or years) and locations (segments or intersections). Most of these studies attempt to explore the relationship between crash counts and explanatory variables, such as roadway geometry, traffic control facilities, traffic conditions, and so on by using Poisson or Negative Binomial regression models (e.g. Poch and Mannering, 1996; Milton and Mannering, 1998; Ivan *et al.*, 1999; Abdel-Aty and Radwan, 2000; Greibe, 2003; Abdel-Aty and Pande, 2007; Wong *et al.*, 2007). For the collective approach, however, individual contributing factors to the crash (e.g., driver demographics, driver behaviors, vehicle types) are not considered and

factors affecting the crash severity cannot be identified either. Therefore, some studies employed individual approach to crash data analysis. The individual approach is characterized by each individual crash case. The main focus of these studies was to associate the crash severity with driver, vehicle and roadway factors by using ordered probit/logit model or logistic regression (e.g., Shanker and Mannering, 1996; Dissanayake *et al.*, 2002; Al-Ghamdi, 2002; Delen, *et al*., 2002; Tay and Rifaat, 2007; Sze and Wong, 2007). More advanced logit-based approaches, such as nested logit model or mixed logit model, were also employed to analyze the same issue (*e.g.* Shanker, *et al.*, 1996; Chang and Mannering, 1999; Milton, *et al.*, 2008).

Although statistic models are the commonly used methods in the context of crash data analysis either collectively or individually, most of them have their own assumptions and complexity in the model estimation and interpretation. Once the assumptions were violated, the model could lead to erroneous estimation results, especially for the individual approach wherein most variables explaining the individual crashes are categorical (*e.g.*, driver gender, road type, lighting condition, violation, weather condition, and severity degree, among others). It is difficult to develop parametric statistical models based upon the categorical data. Therefore, a number of 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), were adopted to deal with the classification and prediction problems. However, two gaps still remain. First, the interpretations of classification results with such methods are weak. The knowledge lying in the crash data cannot be fully discovered, because artificial neural network is in essence a black box and the prediction error of decision tree is usually high. Second, most of statistical methods only provide calibrated parameters with significance tests, which are then used to examine the effects of the corresponding variables on crash counts or crash severity. The interrelationship among explanatory factors cannot be examined in details. According to “error chain theory,” a crash is often caused by a series of errors, not solely by a single factor. As such, mining the explanatory rules is deemed necessary for crash data analysis.

Rule mining, also known as rule generation, rule recovery, or classification/association rule mining, is one of data mining techniques intended to mine for knowledge from available databases and toward decision support. Rule mining is naturally modeled as multi-objective problems with three criteria: (1) predictive accuracy, (2) comprehensibility, and (3) interestingness (Freitas, 1999; Ghosh and Nath, 2004). To automatically search for the optimal combination of rules from a considerable number of potential rules, genetic algorithms (GAs) are perhaps the most commonly used method. By employing GAs to learn of rules is named as 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 performances of rule mining algorithms have been proven and applied in many fields. Thus, this paper aims to develop GMR model that can determine the optimal combination of appraisal rules to achieve the following goals: (1) to discover the key rules that determine the combination of contributing factors’ level to crash severity; (2) to provide the possibility of post-adjustment (fine-tune) of the rules mined; (3) to accurately predict the crash severity. Previous relevant studies have seldom considered the problem of conflict and redundancy among the rules mined, our proposed GMR model will account for the conflict and redundancy in addition to conventional objectives: coverage ratio and predictive accuracy.

The rest of this paper is organized as follows. Section 2 presents the crash data with the definitions of potential contributing factors. The proposed GMR model is narrated in Section 3. Section 4 compares the performance of proposed GMR model with decision tree model. Besides, an in-depth investigation on the mined rules is also discussed. Finally, the concluding remarks and suggestions for future studies follow.

2. DATA

The crash data were collected from 2003-2007 National Traffic Accident Investigation Reports compiled by National Police Agency, Taiwan. Each accident investigation report has been digitized and maintained in the database from which detailed individual crash data of freeway accidents are obtained. The individual crash data include detailed information regarding injury severity of each involved individual, time of accident, driver demographics (age, gender, driver sobriety), involved vehicle types, roadway geometry, traffic control condition, weather condition (clear, rain, fog), pavement conditions (wet, dry), lighting condition, vehicle actions (moving straight, right-turn, left-turn, lane-change), and collision types.

There are 52,117 crash cases occurring on Taiwan's freeways from 2003 to 2007. The injury severity of crashes is determined according to the injury degree of the worst-injured victims in the accident. After screening out incomplete police investigation report, a total of 45,744 crashes are used for this study. Table 1 presents the definition and description of potential explanatory variables to crash severity.

Table 1 Crash data summarized from police accident investigation reports

Information	Variable	Type	Description
Surface condition	x_1	Categorical	1, dry; 2, wet or slippery
Signal control	x_2	Categorical	1, none; 2, yes
Driver gender	x_3	Categorical	1, male; 2, female
Weather	x_4	Categorical	1, sunny; 2, cloudy; 3, rain, storm, fog, etc.
Obstacle	x_5	Categorical	1, none; 2, work zone; 3, others
Lighting condition	x_6	Categorical	1, daytime; 2, dawn or dusk; 3, nighttime with illumination; 4, nighttime without illumination
Speed limit	x_7	Categorical (discretized)	1, 110 KPH; 2, 100KPH; 3, 90-70KPH; 4, 60-40KPH
Road status	x_8	Categorical	1, straight road; 2, grade and curved road; 3, tunnel, bridge, culvert, overpass; 4, others
Marking	x_9	Categorical	1, lane line with marker; 2, lane line without marker; 3, no lane-changing line; 4, no lane line
Use of safety belt	x_{10}	Categorical	1, safety belt fastened; 2, safety belt not fastened; 3, unknown; 4, others
Use of cell phone	x_{11}	Categorical	1, use; 2, not in use; 3, unknown; 4, not driver
License	x_{12}	Categorical	1, with license; 2, without license; 3, unknown
Driver occupation	x_{13}	Categorical	1, in job; 2, student; 3, jobless; 4, unknown
Driver age	x_{14}	Categorical (discretized)	1, under 30 years old; 2, 30-40 years old; 3, 40-50 years old; 4, 50-65 years old; 5, above 65 years old
Travel period	x_{15}	Categorical (discretized)	1, 07:01-09:00 morning peak; 2, 09:01-16:00 day off-peak; 3, 16:01-19:00 afternoon peak; 4, 19:01-23:00 night-peak; 5, 23:01-07:00 midnight to morning
Location	x_{16}	Categorical	1, fast lane, general lane; 2, shoulder, edge; 3, median; 4, accelerating or decelerating lane, ramp; 5, toll plaza and others
Vehicle type	x_{17}	Categorical	1, passenger car; 2, truck; 3, bus; 4, heavy truck, trailer truck, tractor; 5, others
Action	x_{18}	Categorical	1, forward; 2, left lane-change; 3, right lane-change; 4, urgent deceleration or stop; 5, others

Alcoholic use	x_{19}	Categorical	1, no; 2, under 0.25 mg/l (or 0.05%); 3, over 0.25 mg/l (or 0.05%); 4, cannot be tested; 5, unknown
Journey purpose	x_{20}	Categorical	1, work trip or school trip; 2, business trip; 3, transportation activity; 4, visiting, shopping; 5, others or unknown
Major cause	x_{21}	Categorical	1, improper lane-change; 2, speeding; 3, fail to keep a safe distance; 4, alcoholic use; 5, fail to pay attention to the front; 6, other driver's liability; 7, factors not attributed to drivers
Collision type	x_{22}	Categorical	1, head-on or rear-end; 2, sideswipe (common direction); 3, angle or other crash; 4, single-car collision with fixed object; 5, other single-car crash; 6, collision with pedestrian
Severity	y	Categorical	1, fatality; 2, injury; 3, no-injury

In Taiwan, crash severity in police investigation report is classified into three degrees: A1 (fatal crash), A2 (injury crash), and A3 (non-injury crash). The cases for these three degrees of crash severity are 494, 4,073, and 41,177, respectively—an uneven distribution commonly seen in the context of crash analysis. To avoid misleading results caused by sample disproportionate problem, A2 and A3 crash cases are randomly re-selected to the same number as A1 crash cases (494), thus making a total of 1,482 crash cases for our analysis. Furthermore, 70% of these 1,482 crash cases are randomly chosen for training (i.e., 1,037 cases) and the remaining 445 cases are used for model validation. χ^2-test is performed and the result shows that severity distributions between training and validation datasets do not significantly differ.

3. GENETIC MINING RULE MODEL

Genetic mining rule (GMR), which can automatically learn of comprehensive rules from available dataset and toward decision support, is useful in accident analysis (Clarke *et al.*, 1998). The encoding method, fitness function, genetic operators, and rule selection of the proposed GMR model are narrated below.

3.1 Encoding Method

To represent the relationship between explanatory variables and crash severity, each chromosome is used to represent a potential if-then rule. The conditions associated in the "if part" are termed as antecedence part and those in the "then part" are named as consequent part. Besides, the antecedent part consists of at least one variable, but at most 22 variables, selected from Table 2. And the consequent part is composed by, of course, only one variable: severity degree. In general, a rule is 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 degree. Thus, a typical rule i can be of the form: Rule i: If $x_1=a_{i1}$ and $x_2=a_{i2}$ …and $x_j=a_{ij}$ … and $x_{22}=a_{i22}$ Then $y=g_i$. Or, in a shorter form: Rule i: If A_i Then C_i, where a_{ij} is the categorical value of j^{th} attribute variable in rule i. g_i is the value of classification variable in rule i, which ranges from 1 to 3 representing three degrees of crash severity. A_i and C_i are the sets of parties satisfying the antecedent part and consequent part of rule i, respectively.

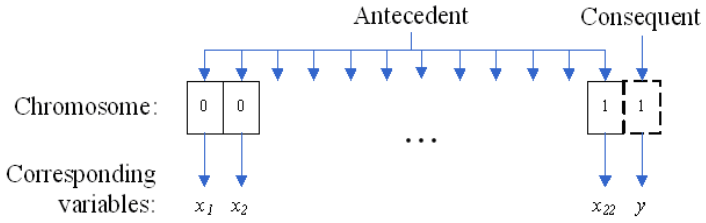

Figure 1 Encoding method of the proposed GMR model

By encoding a rule as a chromosome, each gene is used to represent a corresponding variable. Since the number of potential variables of antecedent and consequent is respectively 22 and one, the length of a chromosome is 23. Each gene will then take one of the categorical values of the corresponding variable. Because the ranges of all variables are different, the ranges of gene values also vary. Moreover, if a gene in a rule antecedent takes a value of 0, it represents the corresponding variable not considered by the rule. If all genes representing the rule antecedent simultaneously take 0 or if the gene representing the rule consequent is 0, then the rule is not included.

Based on this, a rule of "If surface condition=dry and occupation=in job and actions=left lane-change and Then degree of severity=injury" can be encoded as 10000000000010000200002. This rule also contains a family of 2.939×10^{13} offspring rules in total, which can be represented by "If x_1=1 and x_2={0, 1, 2} and x_3={0, 1, 2} and x_4={0, 1,…, 3} and x_5={0, 1,…, 3} and x_6={0, 1,…, 4} and x_7={0, 1,…, 4} and x_8={0, 1,…, 4} and x_9={0, 1,…, 4} and x_{10}={0, 1,…, 4} and x_{11}={0, 1,…, 4} and x_{12}=1 x_9={0, 1,…, 4} and x_{13}=1 and x_{14}={0, 1,…, 5} and x_{15}={0, 1,…, 5} and x_{16}={0, 1,…, 5} and x_{17}={0, 1,…, 5} and x_{18}=2 and x_{19}={0, 1,…, 5} and x_{20}={0, 1,…, 5} and x_{21}={0, 1,…, 7} and x_{22}={0, 1,…, 6} and Then y=2." That is, any case satisfying any one of the offspring rules will certainly also satisfy their parent rule. Generally, the more variable present in the antecedent part (taking non-zero values), the more specific of a rule is (less number of parties will satisfy the rule).

The proposed algorithm aims to select a set of rules which can best predict the severity degree based upon these twenty two explanatory variables. The total number of potential rules equals $3\times3\times4\times4\times4\times5\times5\times5\times5\times5\times5\times5\times5\times6\times6\times6\times6\times6\times6\times6\times8\times7\times4=1.058\times10^{16}$. Obviously, it is barely possible to compare all rule combinations through a total enumeration approach.

3.2 Fitness Function

An individual chromosome, a rule, with a higher fitness function value has a higher probability to be selected for reproducing offspring. The role of fitness function is to evaluate the quality of the rule numerically. In this study we use the following three common factors: coverage, completeness and confidence of the rule. The coverage ratio of rule *i* (*i.e.,* the cases satisfied by the rule antecedent) is denoted by $|A|$: the cardinality of set A (the number of elements in set A). The completeness of the rule (*i.e.,* the proportion of cases of the target class covered by the rule) is given by $|A\cap C|/|C|$. The confidence of rule *i (i.e.,* the predictive accuracy) is given by $|A\cap C|/|A|$ (Freitas, 1999). After several trials on the combination of these three indices, this paper uses predictive accuracy (PA_i) and coverage ratio (CR_i) as the

fitness function (f_i) of rule i, which can be expressed as follows:

$$f_i = 1000 \cdot (CR_i)\ (PA_i)^2 \tag{1}$$

3.3 Genetic Operators

Because the genes in our GMR model are not encoded binary, simple genetic algorithms proposed by Goldberg (1989) cannot be used. Instead, we employ the max-min-arithmetical crossover proposed by Herrera *et al.* (1998) and the non-uniform mutation proposed by Michalewicz (1992). A brief description is given below.

(1) Max-min-arithmetical crossover
Let $G_w^t = \{ g_{w1}^t, \ldots, g_{wk}^t, \ldots, g_{wK}^t \}$ 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:

$$G_1^{t+1} = aG_w^t + (1\text{-}a)G_v^t \tag{2}$$
$$G_2^{t+1} = aG_v^t + (1\text{-}a)G_w^t \tag{3}$$
$$G_3^{t+1} \text{ with } g_{3k}^{t+1} = min\{g_{wk}^t, g_{vk}^t\} \tag{4}$$
$$G_4^{t+1} \text{ with } g_{4k}^{t+1} = max\{g_{wk}^t, g_{vk}^t\} \tag{5}$$

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:

$$g_k^{t+1} = \begin{cases} g_k^t + \Delta(t, g_k^u - g_k^t) & if \quad b = 0 \\ g_k^t - \Delta(t, g_k^t - g_k^l) & if \quad b = 1 \end{cases} \tag{6}$$

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 to 0 as t increases:

$$\Delta(t, z) = z(1 - r^{(1-t/T)^h}) \tag{7}$$

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. (7), the value returned by $\Delta(t, z)$ will gradually decrease as the evolution progresses.

3.4 Rule Selection

To avoid selecting redundant rules during the evaluation process, the rules are selected according to the following steps.

Step 1: Rank rules in the final population (*i.e.* final potential rule set, *FR*) according to their fitness values in a descending order.

Step 2: Select the rule with the highest fitness value from the set of *FR*.

Step 3: Check the redundancy of the rule selected by Step 2. If its redundancy ratio (r_i) is less than 0.5, then remove the rule from *FR* to the mined rule set (*MR*); otherwise, delete the rule from *FR*. The redundancy ratio (r_i) can be expressed as:

$$r_i = \max_{\forall j \in MR} \{ w_{ij} = \frac{D_{ij}}{R_i} \} \tag{8}$$

where, w_{ij} is the degree of redundancy between rules i and j. D_{ij} is the total number of variables in the antecedent part sharing the same value, except for zero, in both rules i

and j. R_i is the total number of variables appearing in the antecedent part of rules i. MR is the set of mined rules.

Step 4: Terminate if no rule left in FR. MR is the optimal combination of rules. Otherwise, go to Step 2.

Even if the chromosomes have been filtered by redundancy index, it is almost inevitable that two or more rules with different predicted classes may be simultaneously fired by a crash case. To synthesize the predicted degree of more than one rules fired, we take an average value of predicted degrees of all fired rules and round it to the nearest integer, which can be expressed as:

$$sg = Int(\frac{1}{|F|}\sum_{j \in F} g_j) \tag{9}$$

where, G is the predicted severity degree by the proposed algorithm. $Int(\cdot)$ is a rounding operator, which rounds value in parenthesis to the nearest integer. F is a set of sequence number of fired rules. As such, the correct rate of the model can be computed as the number of correctly predicted cases divided by the total number of cases.

4. RESULTS

The parameters of the proposed GMR model are set as follows: population size=50, crossover rate=0.85, mutation rate=0.08, and maximum number of generations=1,000. The learning process of the GMR model is shown in Figure 2.

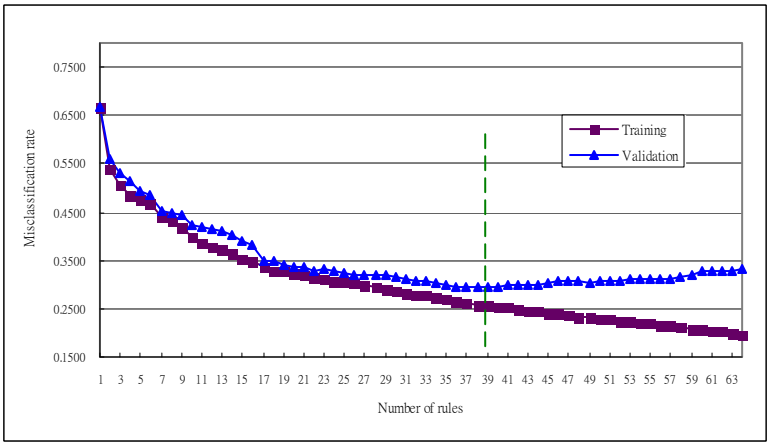

Figure 2 Learning process of the GMR model

Table 2 shows the final selected rules along with its corresponding performance indices. Note that a total of 39 rules are selected with a descending order according to f_i. In terms of fitness value (f_i), the top nineteen rules have remarkably higher values than the rest of twenty rules, suggesting that it is promising to use only the top nineteen rules to conduct the prediction. In terms of coverage ratio (CR_i), R1 can explain 920 cases, followed by R5 (287 cases) and R2 (135 cases). In contrast, some rules cover only very few cases, such as R39 (2 cases) and R36 or R37 (5 cases). In terms of predictive accuracy (PA_i), R26 has the highest predictive accuracy (1.000), followed by R2 (0.941), and R35 the least (0.333).

Table 2 Combination of rules mined by GMR model based on balanced dataset

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}	x_{22}	y	f_i	CR_i	PA_i
R1										1													3	124.76	920	0.375
R2											3												1	115.21	135	0.941
R3																	1		1	5	3	1	3	78.110	100	0.900
R4																			4				1	40.965	52	0.904
R5					1										5								2	38.469	287	0.373
R6																2							1	35.642	76	0.697
R7		1											1								2		1	31.207	47	0.830
R8	1																	5					1	28.302	63	0.683
R9				1												4							3	26.557	102	0.520
R10	1					1	1																2	24.896	93	0.527
R11																1					4		2	23.935	89	0.528
R12		1											4		5				3				1	21.601	35	0.800
R13	1									2													1	21.363	26	0.923
R14							1									1	1					4	2	20.152	49	0.653
R15				1			3								5								1	20.146	46	0.674
R16																						6	1	16.771	23	0.870
R17	1																				6	4	2	15.341	77	0.455
R18																			1			5	2	14.964	58	0.517
R19			1									2											1	14.016	43	0.581
R20														1	5					3			1	9.948	19	0.737
R21						3												3					2	9.143	27	0.593
R22																				3		5	1	6.312	22	0.545
R23				1									1					1	1		5		2	6.249	50	0.360
R24							1															5	2	6.036	27	0.481
R25			1					1						1			1				1		2	6.027	16	0.625
R26	2										1			1					1		2		2	5.786	6	1.000
R27														3			4			3		1	2	5.620	29	0.448
R28													1		2					5	7		2	5.611	11	0.727
R29																		3			6		2	5.143	12	0.667
R30															4					3			1	4.822	20	0.500
R31	1						3							1	5								2	3.635	13	0.538
R32			1										3	1									2	3.444	7	0.714
R33							2			1					3							2	2	3.444	7	0.714
R34	2							2															2	2.679	9	0.556
R35		1						2											1				2	2.250	21	0.333
R36									1						4				2				2	1.736	5	0.600
R37											2												2	1.736	5	0.600
R38				2											4								2	1.286	12	0.333
R39							2													3		5	1	0.482	2	0.500
m	7	3	3	4	1	2	7	3	1	3	3	1	5	6	10	4	4	4	8	7	9	10	-	-	-	-
n	2	0	0	1	0	1	4	2	0	1	2	1	2	1	10	2	1	3	3	7	8	8	-	-	-	-

Note: *m* is the number of variable presence in the selected 39 rules and *n* is the number of variable presence in the selected 39 rules with values not equal to 1.

The importance of variable can be identified by the number of its presence in all rules. The number of variables with values other than 0 (*i.e.* the variable is not considered by the rule) or 1 (*i.e.* the variable describes a normal condition which is meaningless for accident prevention) in all rules is then calculated. In this regard, x_{15} (travel period) is the most important variable which appears in 10 rules, followed by x_{21} (major cause), x_{22} (collision type), and x_{20} (journey purpose). Three variables are shown in only one rule, which are x_5 (obstacle), x_9 (marking), and x_{12} (license), indicating their least significance to crash severity. There are fourteen rules associated with A1 crash, twenty-two rules with A2 crash, and three rules with A3 crash. Although not all the rules can be directly and clearly explained, yet some important information can still be mined from these selected rules. Taking R1 for instance, the rule indicates that once safety belt is fastened, the accident tends to be less severe (A3 crash). As to R6, the accident occurring at shoulder or edge tends to be an A1 crash. In contrast to R6,

R9 indicates that the accident occurring in the accelerating/decelerating lane or ramp tends to be an A3 crash. As to R12, a male driver without a license tends to cause A1 crash. According to R30, the accident occurring during night-peak tends to be an A1 crash. The above-rule interpretations might be useful references for law enforcement or management by the related authorities.

Table 3 gives the distribution of cases with degree of severity predicted by GMR model and with real degree of severity. As shown in Table 3, in the training dataset, the proposed GMR model can actually predict the A3 crash (correct rate 84.68%), followed by A1 crash (75.43%) and A2 (62.61%). The overall correct rate of the proposed GMR model in training has achieved 74.25%. In the validation dataset, the overall correct rate has achieved 70.79%.

Table 3 Number of cases with degree of severity predicted by GMR based on balanced dataset

Datasets	Real severity	Predicted severity			Total
		A1	A2	A3	
Training	A1	261 (75.43%)	58 (16.76%)	27 (7.80%)	346 (100.00)
	A2	58 (16.81%)	216 (62.61%)	40 (20.58%)	345 (100.00)
	A3	13 (3.76%)	40 (11.56%)	293 (84.68%)	346 (100.00)
	Total	332	314	391	1,037
Validation	A1	106 (71.62%)	31 (20.95%)	11 (7.43%)	148 (100.00)
	A2	23 (15.44%)	91 (61.07%)	35 (23.49%)	149 (100.00)
	A3	8 (5.41%)	22 (14.86%)	118 (79.73%)	148 (100.00)
	Total	137	144	164	445

Note: The percentages are given in the parentheses.

5. COMPARISON

For comparison purpose, a decision tree (DT) model is also used to mine the rules explaining the same crash dataset. The DT model is performed by SAS Enterprise Miner Release 4.3. Several settings of the DT model are tried and the best performed settings are 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 learning process of the DT model is depicted in Figure 2. Note that the misclassification rate decreases as the number of leaves gets larger.

Table 4 presents the number of cases with various degrees of severity predicted by the DT model. Note that the DT model performs slightly better in predicting the A2 crash (correct rates in training and validation are 78.84% and 77.18%, respectively) than the proposed GMR model. However, the DT model performs much worse than the proposed GMR model while predicting both A1 and A2 crashes. Averagely, the overall correct rates of the DT model in training and validation are 63.84% and 61.35%, respectively, which are inferior to the proposed GMR model.

In addition, to investigate and compare the performance of the tuned GMR and DT models, all accident cases (a total of 45,744 cases) are used for validation. Results are shown in Table

5. The overall correct rates of the GMR and DT models are 69.62% and 59.24%, respectively, which are slightly lower than those on the balanced dataset. Thus, the performance and applicability of the proposed GMR model has been demonstrated.

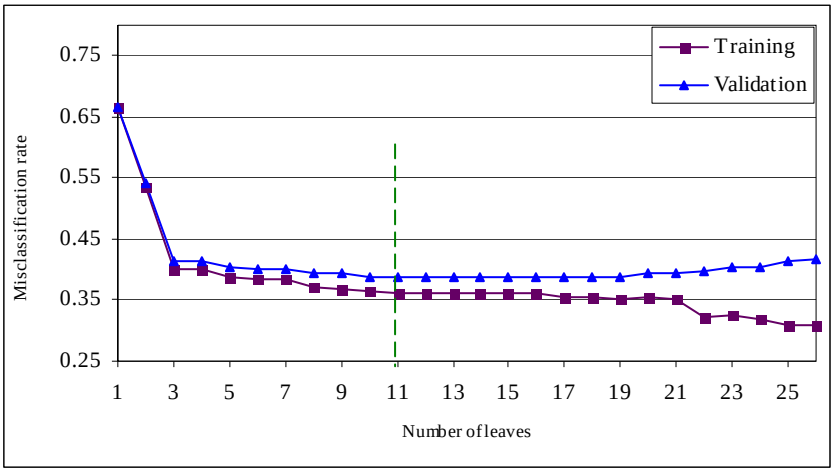

Figure 2 Learning process of the DT model

Table 4 Number of cases with degree of severity predicted by DT based on balanced dataset

Datasets	Real severity	Predicted severity			Total
		A1	A2	A3	
Training	A1	188 (54.34%)	150 (43.35%)	8 (2.31%)	346 (100.00)
	A2	33 (9.57%)	272 (78.84%)	71 (11.59%)	345 (100.00)
	A3	6 (1.73%)	138 (39.88%)	202 (58.38%)	346 (100.00)
	Total	227	560	250	1037
Validation	A1	68 (45.95%)	72 (48.65%)	8 (5.41%)	148 (100.00)
	A2	11 (7.38%)	115 (77.18%)	23 (15.44%)	149 (100.00)
	A3	3 (2.03%)	55 (37.16%)	90 (60.81%)	148 (100.00)
	Total	82	242	121	445

Note: The percentages are given in the parentheses.

Table 5 Comparison of GMR and DT models based on the entire dataset

Model	Real severity	Predicted severity				Total
		A1	A2	A3	Not triggered	
GMR (Overall correct rate = 69.62%)	A1	376 (76.11%)	76 (15.38%)	41 (8.30%)	1 (0.20%)	494 (100.00)
	A2	659 (16.18%)	2,522 (61.92%)	887 (21.78%)	5 (0.12%)	4,073 (100.00)
	A3	3,642 (8.84%)	8,573 (20.82%)	28,951 (70.31%)	11 (0.03%)	41,177 (100.00)
	Total	4,675	10,867	30,185	17	45,744
DT (Overall correct rate = 59.24%)	A1	256 (51.82%)	222 (44.94%)	16 (3.24%)	0 (0.00%)	494 (100.00)
	A2	485 (11.91%)	2,920 (71.69%)	662 (16.25%)	6 (0.15%)	4,073 (100.00)
	A3	1,028 (2.50%)	16,224 (39.40%)	23,864 (57.95%)	61 (0.15%)	41,177 (100.00)
	Total	1,769	19,366	24,542	67	45,744

Note: The percentages are given in the parentheses.

A total of 11 rules are generated by the DT model as follows: four rules associated with A1 crash, five rules with A2 crash, and two rules with A3 crash.
R1: If x_{11}= {3, 4} Then y= 1
R2: If x_{22}= {1, 2, 3} and x_{21}= {1, 2, 4, 5, 6, 7} and x_{11}= {1, 2} Then y = 2
R3: If x_{19}= 1 and x_{17}= {1, 2} and x_{21}= 3 and x_{11}= {1, 2} Then y = 3
R4: If x_{19}= {2, 3, 5} and x_{17}= {1, 2} and x_{21}= 3 and x_{11} ={1, 2} Then y = 2
R5: If x_{20}= {1, 5} and x_{17}= {3, 4} and x_{21}= 3 and x_{11}= {1, 2} Then y = 3
R6: If x_{20}= {2, 3, 4} and x_{17}= {3, 4} and x_{21}= 3 and x_{11}= {1, 2} Then y = 2
R7: If x_{22}=6 and x_{21}= {1, 4, 5, 67} and x_{11}= {1, 2} Then y = 1
R8: If x_{15}= {1, 2, 4, 5} and x_{21}=2 and x_{22}= {4, 5, 6} and x_{11}= {1, 2} Then y = 1
R9: If x_{15}= 3 and x_{21}= 2 and x_{22}= {4, 5, 6} and x_{11}= {1, 2} Then y = 2
R10: If x_{16}= {1, 4} and x_{22}= {4, 5} and x_{21}= {1, 4, 5, 6, 7} and x_{11}= {1, 2} Then y = 2
R11: If x_{16}= {2, 3, 5} and x_{22}= {4, 5} and x_{21}= {1, 4, 5, 6, 7} and x_{11}= {1, 2} Then y = 1

6. CONCLUSION

This paper identified contributing factors to crash severity by developing a novel genetic mining rule (GMR) model. To avoid over-mining caused by unevenly distributed cases across degrees of severity, identical numbers of A1-type, A2-type and A3-type of crash cases are drawn from 2003-2007 Taiwan's freeway accidents dataset. A total of 39 rules have been mined which can achieve an overall correct rate of 74.25% in training and 70.79% in validation, respectively. Our proposed GMR model has demonstrated superior to the conventional decision tree (DT) model, which can only achieve an overall correct rate of 63.84% in training and 61.35% in validation, respectively, with the same database. According to the mined rules, x_{15} (travel period), x_{21} (major cause), x_{22} (collision type), and x_{20} (journey purpose) are the four key factors contributing to crash severity. Consequently, attention must be paid to these four factors to ameliorate the traffic safety.

Some directions for future studies can be identified. First, the neighboring traffic condition of the crash is also an important factor to crash severity; however, the police accident investigation report did not record such information. The crash data may be further matched with the traffic database so as to gain more information regarding the contributing factors to crash severity. Second, since the interrelationship among contributing factors can be quite different across the accident types, segmentation of accident dataset according to accident types (*e.g.*, one-vehicle, two-vehicle and multi-vehicle collisions) prior to the model training will definitely enhance the performance of the proposed model as it would provide more focused information. Such a data-segmentation deserves further attempts. Third, in order to lessen the model complexity, various performance indices may be integrated into an overall fitness function; namely, a multi-objective GMR model deserves further elaboration. Last but not least, more comparisons can be made to other commonly used methods (*e.g.*, logistic regression model, ordered Logit model, artificial neural network) to demonstrate the superiority of the proposed model.

ACKNOWLEDGEMENTS

This research is partially granted by National Science Council, Republic of China (NSC 97-2628-E-009-035-MY3). The authors are indebted to anonymous reviewers for their insightful comments and constructive suggestions to improve the quality of our original manuscript.

REFERENCES

Abdel-Aty, M. and Pande, A. (2007) Crash data analysis: Collective vs. individual crash level approach, **Journal of Safety Research 38**, 581-587.

Abdel-Aty, M. and Radwan, A.E. (2000) Modeling traffic accident occurrence and involvement, **Accident Analysis and Prevention 32**, 633-542.

Al-Ghamdi, A.S. (2002) Pedestrian-vehicle crashes and analytical techniques for stratified contingency tables, **Accident Analysis and Prevention 34**, 205-214.

Chang, L.Y. and Mannering, F. (1999) Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents, **Accident Analysis and Prevention 31**, 579-592.

Chang, L.Y. and Chen, W.C. (2005) Data mining of tree-based models to analyze freeway accident frequency, **Journal of Safety Research 36,** 365-375.

Chang, L.Y. and Wang, H.W. (2006) Analysis of traffic injury severity: An application of non-parametric classification tree techniques, **Accident Analysis and Prevention 38**, 1019-1027.

Chen, T.C. and Hsu, T.C. (2006) A GAs-based approach for mining breast cancer pattern, **Expert Systems with Applications 30**, 674-681.

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**, 777-785.

Clarke, D.D., Forsyth, R.S. and 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**, **No. 2**, 223-234.

Dehuri, S. and Mall, R. (2006) Prediction and comprehensible rule discovery using a multi-objective genetic algorithm, **Knowledge-Based System 19**, 413-421.

Delen, D., Sharda, R. and Bessonov, M. (2006) Identifying significant predictors of injury severity in traffic accidents using series of artificial neural networks, **Accident Analysis and Prevention 38**, 434-444.

Dissanayake, S. and Lu, J.J. (2002) Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes, **Accident Analysis and Prevention 34**, 609-618.

Freitas, A.A. (1999). On rule interestingness measures, **Knowledge-Based Systems 12**, 309-315.

Ghosh, A. and Nath, B. (2004) Multi-objective rule mining using genetic algorithms, **Information Sciences** 163, 123-133.

Goldberg, D.E. (1989) **Genetic Algorithms in Search, Optimization, and Machine Learning**, Addison-Wesley, New York.

Greibe, P. (2003) Accident prediction models for urban roads, **Accident Analysis and Prevention 35**, 273-285.

Herrera, F., Lozano, M. and Verdegay, J.L. (1998) A learning process for fuzzy control rules using genetic algorithms, **Fuzzy Sets and Systems 100**, 143-158.

Ivan, J.N., Pasupathy, R.K. and Ossenbruggen, P.J. (1999) Differences in causality factors for single and multi-vehicle crashes on two-lane roads, **Accident Analysis and Prevention 31**, 695-704.

Michalewicz, Z. (1992) **Genetic Algorithms + Data Structures = Evolution Programs**,

Springer, Berlin.
Milton, J., Shankar, V., and Mannering, F. (2008) Highway accident severities and the mixed logit model: An exploratory empirical analysis, **Accident Analysis and Prevention 40**, 260-266.
Poch, M. and Mannering, F. (1996) Negative binomial analysis of intersection, **Journal of Transportation Engineering 12**, 105-113.
Shanker, V. and Mannering, F. (1996) Statistical analysis of accident severity on rural freeways, **Accident Analysis and Prevention 28**, 728-741.
Shanker V., Mannering, F. and Barfield, W., (1996) Statistical analysis of accident severity on rural freeways, **Accident Analysis and Prevention 28**, 391-401.
Shin, K.S. and Lee, Y.J. (2002) A genetic algorithm application in bankruptcy prediction modeling, **Expert Systems with Applications 23**, 321-328.
Sze, N.N. and Wong, S.C. (2007) Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes, **Accident Analysis and Prevention 39**, 1267-1278.
Tay, R. and Rifaat, S.M. (2007) Factors contributing to the severity of intersection crashes, **Journal of Advanced Transportation 41**, 245-265.
Wong, S.C., Sze, N.N. and Li, Y.C. (2007) Contributory factors to traffic crashes at signalized intersections in Hong Kong, **Accident Analysis and Prevention 39**, 1107-1113.