附件一

行政院國家科學委員會補助專題研究計畫 ☑成果報告 □期中進度報告

基因及螞蟻規則探勘模式-以事故分析及事故鑑定為例(III/III)

Developing Genetic and Ant-based Rule Mining Models- Case Studies on Accident Analysis and Appraisal (III/III)

計畫類別：☑ 個別型計畫　　□ 整合型計畫
計畫編號：NSC 97-2628-E-009-035-MY3
執行期間：97 年 8 月 1 日至 100 年 7 月 31 日

計畫主持人：邱裕鈞　　　　　交通大學交通運輸研究所 教授
計畫參與人員：黃彥斐、閻姿慧、傅強、謝志偉交通大學交研所博士班研究生
邱煜勝、鍾佩儒、高筑韻、蕭任谷、魏嘉儀 交通大學交研所碩士班研究生

成果報告類型(依經費核定清單規定繳交)：□精簡報告 ☑完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
☑出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立交通大學

中華民國 100 年 10 月 17 日

一、摘要

1.1 中文摘要

傳統以個體角度進行事故分析之方法，例如，判別分析（discrimination analysis）、羅吉斯迴歸（logistic regression）、次序普羅比（ordered probit）、羅吉特（logit）及混合羅吉特（mixed probit）等模式，大多僅能探討單一危險因素之影響程度。事實上，事故嚴重與否大多係由多項因素同時發生所導致。此一綜合多項因素之危險情況，在統計分析上，甚難加以窮舉分析。基此，本計畫乃於第一年期提出基因規則探勘模式（Genetic rule mining, GRM），可由探勘所得之規則的前半部，判斷何謂危險情況，進而加以避免。惟本研究所提出之 GRM 必須先固定規則數量，再同時進行最佳規則組合之尋優。因此，具有染色體長度過長，尋優效果不佳，以及探勘過多衝突或重覆規則的傾向，進而導致規則難以詮釋，無法提出具體之安全改善策略。

有鑑於此，本計畫第二年期乃提出改良式的基因規則探勘模式（Genetic rule mining, GRM），稱為逐步基因規則探勘模式（Stepwise GRM, SGRM）。SGRM 一次僅挑選使事故嚴重度預測率精確率最高的一條規則，再以此規則為基礎，進行下一條規則之選取，直到精確率無法再改善為止。如此，即可避免選擇規則過多，且相互重覆或矛盾的問題。此外，由於不同類型事故之影響因素與危險情況不一定相同，因此，有必要加以區隔分析。本年度以先以總計 5563 件單車事故（single vehicle accident）為分析基礎。結果顯示，本模式共選擇了 38 條規則，其訓練準確度達 75.1%，而驗證準確度則達 73.8%均遠高於決策樹之預測結果。而影響事故嚴重度之危險情境也加以確認，並研提改善策略。

本計畫第三年期進一步探討及比較本研究所提出之 SGRM、粗略集合（rough set, RS）及次序普羅比（ordered probit model, OP）三種模式在分析不同事故嚴重度之選擇規則與重要解釋變數。結果顯示，駕駛人職業別、事故地點及車輛型式是三個最主要的關鍵因素。最後，本研究進一步將 SGRM 所挑選出的 38 條推理規則的前半部設定為危險情境（risk condition），以虛擬變數表之，後半部則為事故嚴重度，結合次序普羅比進行危險情境之推估與檢定，以了解各種危險情境對事故嚴重度之影響程度。結果顯示，本研究所提出之整合方法（SGMR＋OP）其模式配適度，遠比將所有原始變數作為解釋變數所建構之 OP 模式為高，更可有效辨識、檢定及推估各種危險情境，有效克服以往統計迴歸方法僅能探討單一變數對事故嚴重度的缺點，更符合事故嚴重度係由多個肇因所導致之先驗知識。

關鍵字：事故分析、逐步基因規則探勘、事故嚴重度、次序普羅比。

1.2 Abstract

Conventional individual approach to conduct accident analysis is to associate the crash severity with driver, vehicle and roadway factors by using discrimination analysis, logistic regression, ordered probit, logit and mixed logit models. Although statistic models are the commonly used methods in the context of crash data analysis, 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. It is difficult to develop parametric statistical models based upon the categorical data. In addition, 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. To this end, the first research year of this project has

proposed genetic rule mining models to discover the key rules (i.e. risky conditions). However, since the proposed GRM models simultaneously select the rule combinations under a given upper limit of rule number and tend to mine too many conflict or redundant rules, making the rule interpretation difficult.

Based on this, the second year of this project further propose a stepwise GRM (SGRM) model, which select the optimal one rule at a time and iteratively proceed to select the next best rule based on the selected rules until model performance (accuracy) can't not improved. Since the risky conditions and contributory factors of various types of crashes will significantly vary, the analysis is conducted on each type of accidents separately. Taking single-vehicle accident for instance, a total of 5,563 crashes on Taiwan's freeway network from 2003 to 2007 are collected, where numbers of A1 (fatal crash), A2 (injury crash), and A3 (property damage only crash) are 226, 1,593, and 3,744, respectively - an uneven distribution commonly seen in the context of crash analysis. A total of 38 rules have been mined which can achieve overall correct rates of 75.1% in training and of 73.8% in validation, respectively, much higher than those yield by the decision tree model. Risky conditions along with their corresponding improvement strategies have been identified.

In the third year, three models-genetic mining rule (GMR), rough set (RS) and ordered Probit (OP) are developed to identify the key factors, wherein a factor with high presence rate at all mined-rules is regarded as the key discriminative factor and that with high presence rate at mined-rules associated with A1 (fatality) is regarded as a key risk factor. The results show that the top three discriminative factors are driver occupation, location, and vehicle type; while the top three risk factors are major cause, driver occupation, and driver age. In addition, a two-stage integrated model combining SGMR and OP are developed, the first stage develops a genetic mining rule (GMR) model to identify possible risk conditions which can best explain the degree of severity. The second stage then develops an OP model with minded risk conditions as dummy explanatory variables. It is found that the proposed two-stage OP model is superior to one-stage OP model in terms of likelihood ratio. Based on the results, six most critical risk conditions have been identified, which can serve as useful guides to ameliorate the traffic safety.

Key Words: *Crash analysis, stepwise genetic rule mining, crash severity, ordered Probit model.*

二、主要研究成果

2.1 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. It is shown in Figure 1 that limited information could be mined from the influence of single variable on crash severity. In contrast, combination of multiple variables would reveal explicit tendency in crash severity as shown in Figure 2 (The four rules in it is selected from the final rule set in this study).

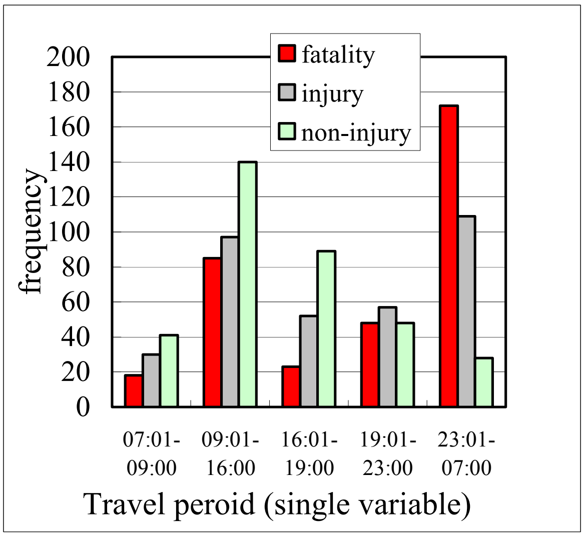

Figure 1 analysis of single variable

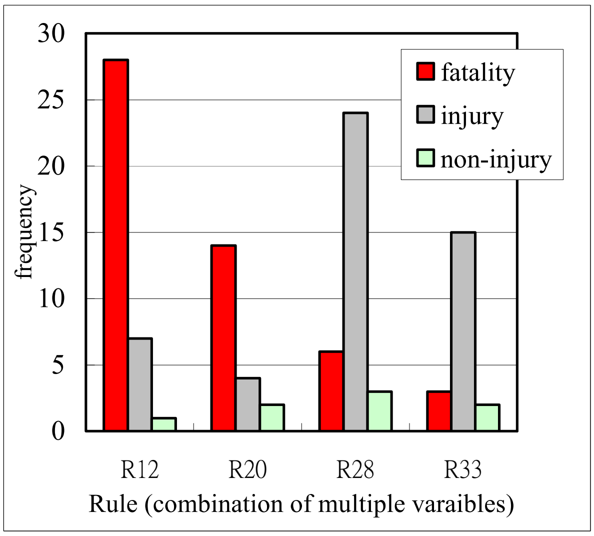

Figure 2 applying rules to analyze

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 decision 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.

2.2 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, and vehicle actions (moving straight, right-turn, left-turn, lane-change).

Considering the characteristics of crash occurrence may differ in collision type, the single-vehicle accident data are chosen to diminish the heterogeneity of crash data. Single-vehicle accidents are those in which only a single vehicle is involved. There are 5,563 single-vehicle 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. 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, others or unknown
Use of cell phone	x_{11}	Categorical	1, use; 2, not in use; 3, others or unknown
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
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 226, 1,593, and 3,744, respectively—an uneven distribution commonly seen in the context of crash analysis. Furthermore, 70% of these 5,563 crash cases are randomly chosen for training (i.e., 3,895 cases) and the remaining 1,668 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.

2.3 Genetic rule mining 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.

2.3.1 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 21 variables, selected from Table 1. 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_{21}=a_{i21} 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.

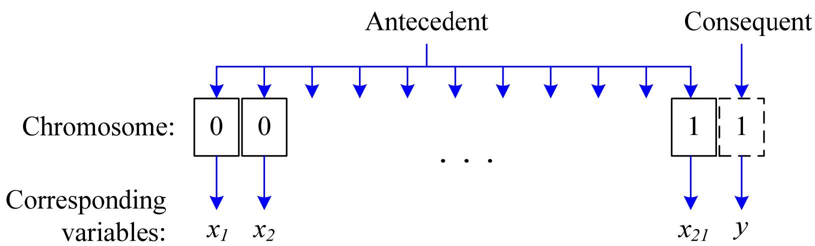

Figure 3 Encoding method of the proposed GMR model

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 21 and one, the length of a chromosome is 22. 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 1000000000001000020002. This rule also contains a family of 4.838×10^{10} 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}={0, 1,…, 3} 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 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 one explanatory variables. The total number of potential rules equals $3\times3\times3\times4\times4\times5\times5\times5\times5\times4\times4\times4\times5\times6\times6\times6\times6\times6\times6\times6\times8=1.935\times10^{14}$. Obviously, it is barely possible to compare all rule combinations through a total enumeration approach.

2.3.2 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. To determine the fitness function, there are three common factors frequently taken into consideration: 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). Shin and Lee(2002) adopted hit ratio(confidence) as the fitness function which is also defined as predictive accuracy plus coverage in another study(Kim and Han, 2003). However, it is the performance of the entire rule set that should be emphasized instead of those ones of individual rules themselves. In other words, the good performances of individual rules do not guarantee that the combination of these rules also performs well. It results from the redundancy and conflict between rules. In order to overcome this problem, the fitness function is set in this paper as the increase of correctly classified cases by the rule set combining the previous mined rules and the new rule, which can be expressed as follows:

$$f_i = N_{nrs} - N_{prs} \tag{1}$$

where, N_{nrs} is the number of cases that are correctly classified by the rule set combining the previous mined rules with the rule *i*, and N_{prs} is the number of cases that are correctly classified by the previous mined rules.
The previous mined rules are also called the temporary rule set in this study. By means of the fitness function above, the effect caused by redundancy or conflict between rules would be effectively reduced in rule mining process. When a new rule is extracted from the final population, it would certainly increase the performance of entire rule set as the new rule set combines the new rule with the temporary rule set.

2.3.3 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 ,…, g_{wk}^t ,…, g_{wK}^t } and G_v^t ={ g_{v1}^t ,…, g_{vk}^t ,…, g_{vK}^t } be two chromosomes selected for crossover, the following four offsprings can be generated:

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

(2)

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

(3)

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

(4)

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

(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:

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

(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:

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

(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.

2.3.4 Rule selection

The method of extracting rules has profound effects on their accompanied performance. Conventionally, a group of different rules is obtained simultaneously from the final results as the stopping criterion is met. Generally speaking, it is an important issue to avoid selecting redundant or conflicting rules during the rule selection process. The redundancy or conflict between the selected rules would lead to reduce the performance of the prediction model, as well as increasing the difficulty in interpreting the causal relationship between explanatory variables and crash severity. However, it is probably difficult to avoid this condition and little information could be found in the literature on dealing with this issue (Shin and Lee, 2002; Kim and Han, 2003; Chen and Hsu, 2006). On the other hand, the mined rules are often too complicated to be understood instead of being interpretable, shorter, and simpler. In order to improve these problems, a learn-one-rule function combining with a neighborhood search was introduced over the rule mining process in this study. Instead of searching a good rule set at a time, a stepwise rue set building procedure with a greedy strategy is proposed. Applying the learn-one-rule function combining with a neighborhood search, the rule set is constructed according to the following steps (as shown in Figure 4):

Step 1: Rank rules in the final population according to their fitness values in a descending order.
Step 2: Select the rule with the highest fitness value and perform a neighborhood search with improvement and parsimony principle for rule modification.
Step 3: Update the temporary rule set by the modified rule.
Step 4: Terminate until the number of rules in the temporary rule set hit the preset number. Otherwise, implement the GAs for another run and go to Step 1.

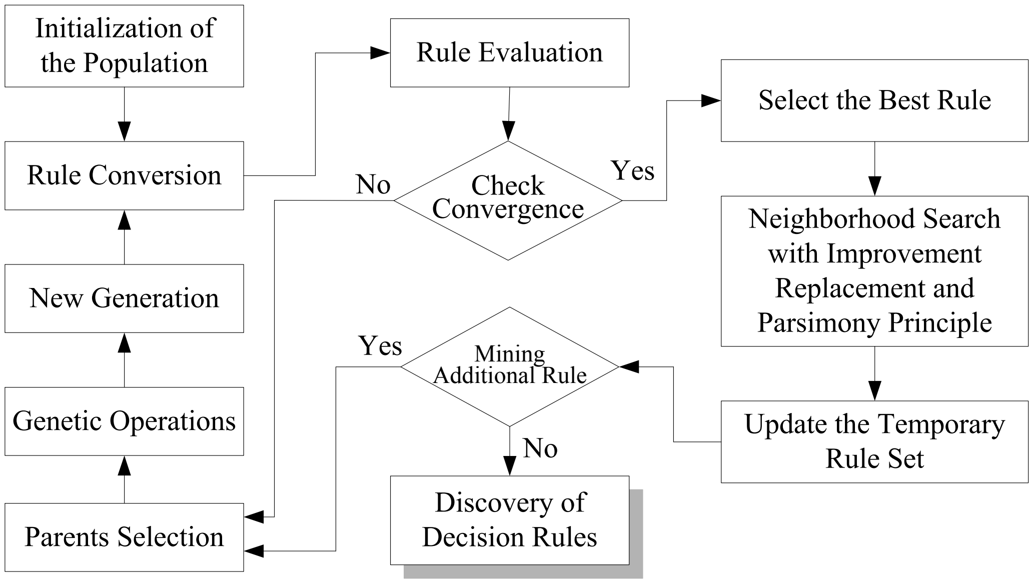

Figure 4 The GA based mining approach

After a rule is selected, a rule modification scheme is introduced. There are two mechanisms in the rule modification process, including improvement replacement and parsimony principle. Due to the characteristic of stochastic operation in evolutional process, it is understandable that there might be some better points existing near the current solution point in the search space. Based on this, Comparative rules are created by enumerating all other attribute values of one variable controlling all other variables. In the mechanism of improvement replacement, when the predictive accuracy of a comparative rule combining with the previous rule set is better than the raw rule in the same condition, the value of the checked variable would be substituted by the value of the same variable in that comparative rule, as shown in the left part in Figure 5. If there is no better point found, the mechanism of parsimony principle will hold. When the original value of the checked variable is not zero, but the value of the checked variable is zero in comparative rule with the same predictive accuracy in the same condition, the value of the checked variable would be substituted by zero, as shown in the right part in Figure 5. In this study, the order of checking all explanatory variables is from x_1 to x_{21}. After all explanatory variables are checked, the last adjusted rule will be put into the temporary rule set for next rule mining if needed.

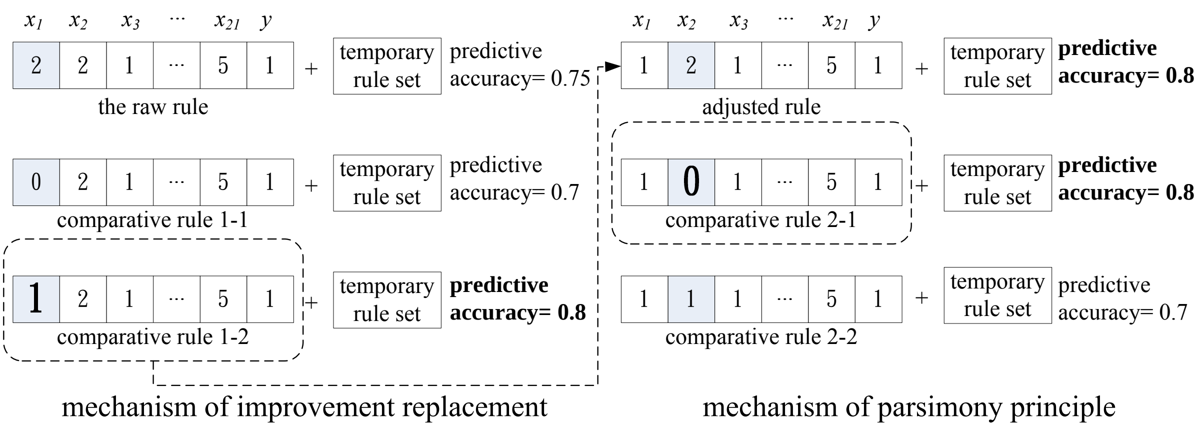

Figure 5 Rule modification process

It is almost inevitable that two or more rules with different predicted classes may be simultaneously fired by a crash case. In this situation, the case is would be predicted as the class of the rule with the highest accuracy if two or more rules are applied to the case at the same time.

2.4 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 stopping criterion). The number of rules to mine is set as 55. The learning process of the GMR model is shown in Figure 6.

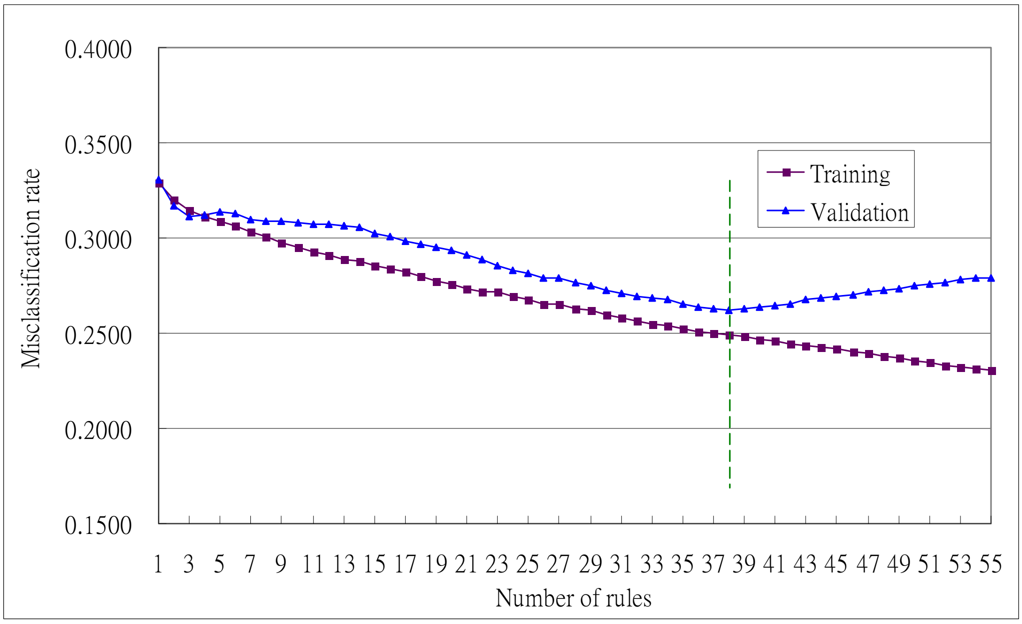

Figure 6 Learning process of the GMR model

Theoretically, the misclassification rate can be lowered to zero monotonically by increasing the number of rules in the GMR model. However, a good classification model should not only fit the training data well, it must also accurately classify records it has never seen before. To avoid model overfitting, 38 rules are selected in the GMR model as the misclassification rate of validation data hit the lowest value. Table 2 shows the final selected rules along with its corresponding performance indices. Note that a total of 38 rules are selected with a descending order according to PA_i. In terms of predictive accuracy (PA_i), the top twenty five rules have remarkably higher values than the rest of thirteen rules. In terms of coverage ratio (CR_i), R23 can explain 3,800 cases, followed by R30 (1,460 cases) and R31 (529 cases). In contrast, some rules cover only very few cases, such as R1 (6 cases), R6 (6 cases) or R7 (6 cases).

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) in all rules is then calculated. In this regard, x_{13} (driver occupation) is the most important variable which appears in 16 rules, followed by x_{16} (location), x_{15} (travel period), and x_{17} (vehicle type). Two variables are shown in less than three rules, which are x_2 (signal control) and x_8 (road status), indicating their least significance to crash severity. There are six rules associated with A1 crash, twenty-eight rules with A2 crash, and four rules with A3 crash.

Most of the rules could be readily inspected and explained by the if-then relationship of the rules themselves. Taking R1 for instance, the rule indicates that when speed limit is 40~60 KPH and driver's age is over 65 years old, it tends to lead A2 crash. R2 shows when drivers are male, in job and under 30 years old, speed limit is 100 KPH, travel period is midnight to morning, and major cause is alcoholic, it tends to lead A2 crash. As to R19, when safety belt is not fastened with driver's speeding, it tends to cause A1 crash. In contrast to R19, R23 reveals when safety belt is fastened, it tends to be less severe (A3 crash). The rest may be deduced by analogy. More exploration of the potential implications of the rules is depicted as the following. In regard to driver characteristics, it is interesting that jobless driver combining with specific conditions would tend to cause A2 crash. The conditions include cloud (R3), nighttime with illumination, under 30 years old, and midnight to morning (R20), and no obstacle (R26). Regarding Behavior and environment factors, when safety belt is not fastened with driver's speeding, it tends to cause A1 crash (R19). Use of cell phone combining with the antecedents of R14 and R35 tends to lead A2 crash. The alcoholic use has positive correlation in crash severity. On the other hand, wet or slippery surface condition and obstacle do not have significant effects on crash severity. About vehicle type, truck combining with the antecedents of R6, R13, R18, and R27 is likely to lead A2 crash. As to trip characteristic, midnight to morning combining with the antecedents of R2, R5, R20, R21, and R29 also tends to lead A2 crash. The above-rule interpretations might be useful references for law enforcement or management by the related authorities.

Table 2 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
R1	1					1	4		1					5					1			2	6	1.000
R2			1	1			2			1			1	1	5			1			4	2	12	0.917
R3				2									3									2	12	0.917
R4			2				1					1				1		2				2	10	0.900
R5	1		1												3	1			3			2	12	0.833
R6				1									1		5	2	2					2	6	0.833
R7	1				1			1		3	3				2							2	6	0.833
R8																	5		3			2	12	0.833
R9			1			3			3				1			1		1				2	11	0.818
R10				1			2		1					4				1		5	7	2	16	0.813
R11				1			1					2					1					2	16	0.813
R12						1	3						1							5		3	64	0.813
R13	1											2			2		2					2	10	0.800
R14					1		3				1								2			2	15	0.800
R15			1	1												1	4		1			3	239	0.799
R16	1				1	4							1	3			2					2	22	0.773
R17												1				2			2	5		2	12	0.750
R18			2										1		2		2			5		2	12	0.750
R19										2											2	1	11	0.727
R20						3							3	1	5							2	11	0.727
R21													4		5	1	2					2	25	0.720
R22	1						3		1					2			2					2	14	0.714
R23										1												3	3800	0.687
R24						4								1		1			1			3	201	0.687
R25											3											1	106	0.613
R26					1								3									2	154	0.435
R27				1													2			1		2	77	0.429
R28																			4			1	47	0.426
R29	1												1	3	5			1				2	91	0.374
R30	1				1			1	1			1				1						2	1460	0.325
R31						4			1				1									2	529	0.319
R32							3					1	1			1						2	305	0.302
R33															2			2				2	64	0.297
R34	2														4		1					2	149	0.262
R35						1	1			1	1	1	1			2		1	1			2	121	0.215
R36	1												1								2	1	97	0.196
R37			1	1		3								2		1						1	75	0.080
R38													1			2		1	1			1	267	0.064
m	10	0	7	8	5	9	10	2	6	5	4	7	16	9	11	13	11	8	10	5	4	-	-	-

Note: *m* is the number of variable presence in the selected 38 rules.

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 80.77%), followed by A2 crash (64.90%) and A1 (53.13%). The overall correct rate of the proposed GMR model in training has achieved 75.10%. In the validation dataset, the overall correct rate has achieved 73.80%.

Table 3 Number of cases with degree of severity predicted by GMR

Datasets	Real severity	Predicted severity A1	A2	A3	Total
Training	A1	85 (53.13%)	46 (28.75%)	29 (18.13%)	160 (100.00)
	A2	32 (2.87%)	723 (64.90%)	359 (32.23%)	1114 (100.00)
	A3	22 (0.84%)	482 (18.39%)	2117 (80.77%)	2621 (100.00)
	Total	139	1251	2505	3895
Validation	A1	37 (56.06%)	15 (22.73%)	14 (21.21%)	66 (100.00)
	A2	3 (0.63%)	307 (64.09%)	169 (35.28%)	479 (100.00)
	A3	11 (0.98%)	225 (20.04%)	887 (78.98%)	1123 (100.00)
	Total	51	547	1070	1668

Note: The percentages are given in the parentheses.

2.5 Comparisons

2.5.1 Decision tree (DT)

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 7. 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 better in predicting the A3 crash (correct rates in training and validation are 97.71% and 97.15%, 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 70.24% and 69.54%, respectively, which are inferior to the proposed GMR model.

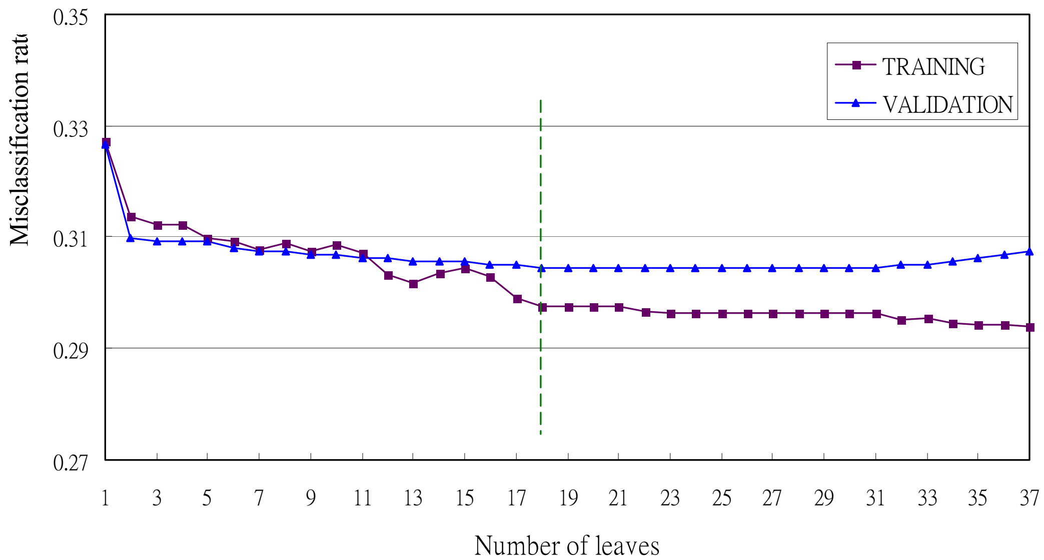

Figure 7 Learning process of the DT model

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	71 (44.38%)	10 (6.25%)	79 (49.38%)	160 (100.00)
	A2	34 (3.05%)	104 (9.34%)	976 (87.61%)	1114 (100.00)
	A3	10 (0.38%)	50 (1.91%)	2561 (97.71%)	2621 (100.00)
	Total	115	164	3616	3895
Validation	A1	36 (54.55%)	1 (1.52%)	29 (43.94%)	66 (100.00)
	A2	7 (1.46%)	33 (6.89%)	439 (91.65%)	479 (100.00)
	A3	7 (0.62%)	25 (2.23%)	1091 (97.15%)	1123 (100.00)
	Total	50	59	1559	1668

Note: The percentages are given in the parentheses.

A total of 18 rules are generated by the DT model as follows: two rules associated with A1 crash, six rules with A2 crash, and ten rules with A3 crash.

R1: If x_{11}=3 Then y=1
R2: If x_{11}=2 Then y=3
R3: If x_{21}=2 and x_{10}= {2, 3} and x_{17}= {1, 4} and x_{11}=1 Then y=1
R4: If x_3=2 and x_4= {2, 3} and x_{17}= {2, 3, 5} and x_{11}=1 Then y=2
R5: If x_3=1 and x_4= {2, 3} and x_{17}= {2, 3, 5} and x_{11}=1 Then y=3
R6: If x_{12}=1 and x_{19}=1 and x_{10}=1 and x_{17}= {1, 4} and x_{11}=1 Then y=3
R7: 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=3
R8: 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=2
R9: 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=3
R10: 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=3
R11: 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=2
R12: 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=3
R13: 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=3
R14: 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=2
R15: 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=2
R16: 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=3
R17: 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=3
R18: 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=2

2.5.2 Rough set model

Rough set (RS) theory, proposed by Pawlak (1982), is an extension of set theory that can effectively handle discrete variables with multilevel categories. The RS theory can classify accidents into groups with similar properties by considering multiple dimensions. It is believed that rough set theory has the potential to be a complementary method for analyzing relationships among factors and crash severity. The RS theory is briefly narrated below.

Let U represent the universe, a finite set of objects, and A denote a set of condition attributes, i.e. affecting factors for crash severity. For x, $y \in U$, we say that x and y are indiscernible by the set of condition attributes A if $\rho(x,a) = \rho(y,a)$ for every $a \in A$ where $\rho(x,a)$ denotes the information function. A set with objects within it which are indiscernible by the set of condition attributes A is called a A-elementary set. The family of all elementary sets is denoted by A*. It represents the smallest partitions of objects by the specified condition attributes so that objects belonging to different elementary sets are discernible and those belonging to the same elementary sets are indiscernible. The A-lower approximation of a set of objects Y ($Y \subseteq U$), denoted by $\underline{AY}$, and the

A-upper approximation of Y, denoted by $\overline{AY}$, are respectively defined as

$$\underline{AY} = \bigcup X \quad \{X \in A^* \quad \text{and} \quad X \subseteq Y\} \tag{8}$$
$$\overline{AY} = \bigcup X \quad \{X \in A^* \quad \text{and} \quad X \cap Y \neq \phi\} \tag{9}$$

The objects belonging to the set of lower approximation are those definitely definable by the elementary sets, since objects in $\underline{AY}$ can be fully identified by the elementary sets in A^*. On the other hand, those belonging to the set of upper approximation but not to the set of lower approximation cannot be fully identified by the elementary sets in A^*.

As illustrated in Table 5, five cases are characterized with three condition attributes: driver's age, vehicle type and weather, and one decision attribute: crash severity. The three condition attributes form four elementary sets: {1, 3}, {2}, {4}, {5}. This means that cases 1 and 3 are indiscernible while the other cases are characterized uniquely with all available information. Therefore, the fatal accident type is described with the lower approximation set {2} and the upper approximation set {1, 2, 3}. Similarly, the concept of the injury accident type is characterized by its lower approximation set {4, 5} and upper approximation set {1, 3, 4, 5}. The performance of the specified condition attributes can be measured with two indicators: accuracy of approximation and quality of approximation. Accuracy of approximation represents the percentages of the associated objects definable with the specified condition attributes. This can be defined as follows:

Table 5 Example of crashes with three factors

Case	Driver occupation	Vehicle type	Weather	Severity
1	In job	Passenger car	Rainy	A1
2	Jobless	Truck	Rainy	A1
3	In job	Passenger car	Rainy	A2
4	Student	Passenger car	Rainy	A3
5	Student	Truck	Sunny	A2

$$\mu_A(Y) = \text{card}(\underline{AY}) \Big/ \text{card}(\overline{AY}) \tag{10}$$

where card refers to cardinality. The accuracy value ranges from 0 to 1. The closer to 1 is the accuracy, the more discernible is the accident type. Namely, more cases of this accident type are discernible by the elementary sets generated by the specified condition attributes. It implies that the associated crash severity do exist unambiguously. Following Table 4, the accuracy of approximation for the fatal class is 0.33(=1/3) and for the injury class is 0.50(=2/4). This implies the injury class can be defined more unambiguously than the fatal class with the provided three condition attributes. On the other hand, quality of approximation represents the definable percentage of the whole universe. Let $X = \{Y_1, Y_2, \ldots, Y_n\}$ be a classification of U, i.e. $Y_i \cap Y_j = \varphi$, $\forall\, i, j \leq n$ $i \neq j$ and $\bigcup_{i=1}^{n} Y_i = U$. Y_i are called classes of X. The A-lower approximation and A-upper approximation of X are represented by sets $\underline{AY}$ = { $\underline{AY_1}$, $\underline{AY_2}$, ... , $\underline{AY_n}$ } and $\overline{AY}$ = { $\overline{AY_1}$, $\overline{AY_2}$, . . . , $\overline{AY_n}$ }, respectively. Quality of approximation of classification X by a set of attributes can be defined as follows:

$$\eta_A(Y) = \bigcup \text{card}(\underline{AY_i}) \Big/ \text{card}(U) \tag{11}$$

The value of quality ranges from 0 to 1. The closer to 1 is the quality, the more objects of the universe clearly belong to a single class of *X,* suggesting that the crash severity for all accidents can be clearly identified. Accidents thus can be more accurately recognized and corresponding countermeasures devised. The quality of approximation for the example is 0.60(=3/5), suggesting that with the provided three condition attributes 60% of the cases can be unambiguously defined. To recognize further the details of crash severity, rules need to be extracted. A rule, representing the critical characteristics of the associated accidents, is a combination of values of condition and decision attributes. Theoretically, the maximum number of rules is the product of the categories of all condition attributes. However, some combinations may not appear since such accidents have never happened before. A rule exists if and only if at least one such accident exists. This paper applies the most frequently used minimum covering algorithm to generate rules, with aims to generate the minimum number as well as the shortest length of rules to cover all accidents.

A rough set software, ROSE2 (rough set data explorer), is used in this paper wherein LEM2 (Grzymala-Busse, 1992; Grzymala-Busse and Werbrouck, 1998) is embedded to generate a minimum rule set covering all cases. The results of RS model are summarized in Table 6. The accuracy of approximation for each class of crash severity is commonly high and the quality of classification is rather low, except for property-damage only crashes. The 5-fold cross-validation technique is used to conduct validation test of classification results.

Table 6 Results of the RS model

Crash severity	Generated rules	Accuracy of approximation (%)	Quality of classification (%)	Hit rate (%)	Overall hit rate (%)
A1		84.40		30.08	
A2	1644	72.96	90.38	27.31	59.01
A3		86.74		74.25	

2.5.3 Ordered Probit model

To further compare the key factors identified by conventional statistical methods, an ordered Probit (OP) model is estimated. The OP model is usually in a latent (i.e., unobserved) variables framework with the following general specification:

$$y_i^* = \beta' X_i + \varepsilon_i \qquad (12)$$

where yi^* is a latent and continuous measure of injury severity faced by an accident victim i; β' is the vector of estimated parameters; X_i is the $(K\times 1)$ vector of observed non-stochastic explanatory variables. What can be observed are:

$$\begin{aligned} &Y^* = 0 \quad \text{if} \quad -\infty < y^* \leqq \mu_0 \quad \text{(property-damage only)} \\ &Y^* = 1 \quad \text{if} \quad \mu_0 < y^* \leqq \mu_1 \quad \text{(injury)} \\ &Y^* = 2 \quad \text{if} \quad \mu_1 < y^* < \infty \quad \text{(fatality)} \end{aligned} \qquad (13)$$

The μ's are unknown threshold parameters to be estimated with β. The method of maximum likelihood is used for estimating the parameters of OP model and the results are presented in Table 7 (only significant factors are listed). Table 8 reveals the number of cases with each severity level predicted by the OP model and the hit rates of training and validation data sets.

Table 7 Estimation results of the OP model

Variable		β	t-value
Constant		0.656	3.89
Thresholds	μ_1	0.00	-
	μ_2	1.430	35.10
Driver characteristics			
License (without license)		0.171	1.75
Use of safety belt (safety belt fastened)		-1.733	-13.92
Driver occupation (student)		0.206	2.66
Alcoholic use (> 0 mg/l)		0.182	2.94
Journey purpose (visiting/shopping trip)		0.151	2.35
Driver age (over 50 years old)		0.506	2.54
Vehicle characteristics			
Vehicle type (truck; bus)		0.232	4.55
Action (others)		0.166	1.74
Crash characteristics			
Location (shoulder edge; median)		0.168	3.04
Major cause (fail to keep a safe distance)		-0.506	-3.54
Travel period (midnight to morning)		0.109	2.25
Environmental characteristics			
Speed limit		0.277	2.15
Road status (grade and curved road)		0.247	1.93
Marking (no lane-changing line)		0.381	2.48
Surface condition (dry)		0.187	4.25
Obstacle (work zone; others)		-0.214	-2.40
Light (nighttime without illumination)		0.117	2.21
Goodness of fit measures			
Mean Loglikelihood (null model)			-0.756
Mean Loglikelihood (convergence)			-0.707
Adjusted rho-square			0.065
BIC			5654.052
AIC			5543.238

Note: The significance (t-value) of independent variables is above 1.645.

Table 8 Number of cases with degree of severity predicted by the OP model

Datasets	Real severity	Predicted severity			Total counts
		A1	A2	A3	
Training (hit rate = 68. 74%)	A1	28 (17.50%)	22 (13.75%)	110 (68.75%)	160
	A2	23 (2.06%)	40 (3.59%)	1051 (94.34%)	1114
	A3	4 (0.15%)	23 (0.88%)	2594 (98.97%)	2621
	Total	55 (1.41%)	85 (2.18%)	3755 (96.41%)	3895
Validation (hit rate = 68.37%)	A1	15 (22.73%)	14 (21.21%)	37 (56.06%)	66
	A2	8 (1.67%)	17 (3.55%)	454 (94.78%)	479
	A3	1 (0.09%)	12 (1.07%)	1110 (98.84%)	1123
	Total	24 (1.44%)	43 (2.58%)	1601 (95.98%)	1668

Note: The percentages are given in the parentheses.

2.6 The proposed two-stage model

This paper proposes a two-stage analytical framework to identify the critical risk conditions contributing to crash severity. The first stage develops a genetic mining rule (GMR) model to identify possible risk conditions best elucidating the degree of severity. The second stage then uses the mined risk conditions as dummy explanatory variables to formulate an OP model. The proposed two-stage analytical framework is applied to analyze the Taiwan's empirical one-vehicle crash data. A total of 38 rules are mined, which can achieve overall prediction rates of 75.10% in training and 73.80% in validation.

The results of two-stage OP model are shown in Table 9. For comparison, a one-stage OP model using the same 21 original explanatory variables is also examined. The estimated results show that the proposed two-stage OP model is superior to one-stage OP model in terms of likelihood ratio ($P < 0.05$).

Table 9 Estimation results of the OP model

Variables	Coefficient	*P*-value	Variables	Coefficient	*P*-value	Variables	Coefficient	*P*-value
Constant	0.3383	0.0055	R14	1.2627	0.0000	R24	0.7274	0.0002
R2	1.7463	0.0000	R15	0.9255	0.0008	R25	1.0629	0.0000
R3	0.9208	0.0000	R16	0.8696	0.0004	R26	0.3404	0.0001
R4	0.4685	0.0000	R17	0.9047	0.0001	R27	0.3241	0.0068
R6	0.3264	0.0004	R18	1.1862	0.0003	R29	0.1852	0.0000
R7	1.2926	0.0022	R19	0.8726	0.0003	R34	-0.3123	0.0182
R8	1.0258	0.0001	R20	0.7978	0.0003	R36	-0.3710	0.0000
R10	0.9913	0.0006	R21	1.0150	0.0008	R37	-1.0080	0.0000
R11	0.8589	0.0007	R22	0.9401	0.0039	Mu(1)	1.5857	0.0000
R12	0.8678	0.0317	R23	0.5951	0.0343			
Log likelihood function			-3721.896					
Restricted log likelihood			-4198.595					
Likelihood Ratio Index				0.114				

Note: The coefficients of R1, R5, R9, R13, R28, R30~R33, R35, R38 are not significant (at $\alpha = 0.1$).

According to the results of OP model with rule-based dummy variables, the risk conditions can be identified. The magnitude of the coefficients in the model implies how severe the corresponding risk conditions will cause the crash severity. Based on the results, special attention should be given to the six most critical risk conditions: R7, R14, R18, R25, R8, and R21 (with coefficients greater than 1). R7 reveals that the elder drivers are in high risk when they pass through the road section with low speed limit (*e.g.*, work zone); thus, more visible signs of variable speed limits or other safer devices might be helpful. For R14, the geometry and signs for freeway on-ramp should be designed to prevent the drunken drivers, motorcyclists or bicyclists from wrongly entering. For R18, the truck drivers without licenses tend to be in high risk during the daytime off-peak periods; thus, more intensive enforcement might need to correct this illegal driving behavior. For R25, due to the characteristics of truck manipulation, it is difficult to cope with the decrease in speed limits; thus, the authority may introduce more appropriate measures or strategies in the speeds control. R8 is a commonly seen problem, which requires more intensive enforcement or education. The light alcoholic use is likely to influence the driving safety. As for R21, this kind of driving behavior must be forbidden. In sum, the above rule interpretations can provide the police with valuable guidance for law enforcement; they can also provide other authorities with helpful information for traffic management.

2.7 Conclusion

This study identifies risky conditions (joint effects of risk factors) to crash severity by developing a novel genetic mining rule (GMR) model. Three different types of A1, A2 and A3 single-vehicle crash cases are drawn from 2003-2007 Taiwan's freeway accidents dataset. A total of 38 rules have been mined which can achieve an overall correct rate of 75.10% in training and 73.80% in validation, respectively. Our proposed GMR model has demonstrated superior to the conventional decision tree (DT), rough set (RS) and ordered Probit (OP) models, which can only achieve an overall correct rate of 70.24% in training and 69.54% in validation, respectively, with the same database. According to the mined rules, x_{13} (driver occupation), x_{16} (location), x_{15} (travel period), and x_{17} (vehicle type) are the four key factors contributing to crash severity. In addition, this study also identified the joint effects of risk conditions on crash severity by developing a novel two-stage model (GMR+OP) model by introducing these 38 mined rules (*i.e.* risk conditions) into an OP model as dummy explanatory variables, it is found that the proposed two-stage OP model is superior to one-stage OP model in terms of likelihood ratio. Based on the results, six most critical risk conditions have been identified, which can serve as useful guides 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, 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. Other information such as the neighboring traffic condition of the crash can also be an important factor to crash severity. Future study can combine such information so as to gain deeper insights into the risk conditions to crash severity. Last but not least, analysis of two-vehicle or multi-vehicle crash data is worthy of further study. Finally, risk conditions containing fewer (only two or three) original explanatory variables would be much easier to elucidate the relationship between explanatory variables and crash severity, thus deserves further attempt.

三、計畫成果自評

國科會補助專題研究計畫成果報告自評表

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

☑達成目標

☐ 未達成目標（請說明，以 100 字為限）

☐ 實驗失敗

☐ 因故實驗中斷

☐ 其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：☑已發表 ☐未發表之文稿 ☐撰寫中 ☐無

專利：☐已獲得 ☐申請中 ☐無

技轉：☐已技轉 ☐洽談中 ☐無

其他：(以 100 字為限）

本計畫發表文章與指導論文寫作臚列如下：

(1)國際研討會論文 4 篇：

Chiou, Y.C., Lan, W.L. and Chen, W.B. (2009) "Contributory factors to crash severity in Taiwan freeways: Genetic mining rule approach," *presented at the 8th International Conference of Eastern Asia Society for Transportation Studies,* Surabaya, Indonesia, Nov. 16-18.

Chiou, Y.C., Lan, W.L. and Chen, W.B. (2010) "Identification of risky conditions contributing to crash severity with genetic mining rules," *presented at the 15th Conference of Hong Kong Society for Transportation Studies,* Hong Kong, China, Dec. 11-14.

Chiou, Y.C., Hwang, C.C. Chang, C.C. and Fu, C. (2011) "A bivariate generalized ordered probit approach for two-party crash severity modeling," *presented at the 3rd International Conference on Road Safety and Simulation*, September 14-16, 2011, Indianapolis, USA. (Invited submission to *Accident Analysis and Prevention*)

Chiou, Y.C., Lan, W.L. and Chen, W.B. (2011) "Exploring key discriminative and risk factors affecting one-vehicle crash severities in Taiwan freeways," *presented at the 9th International Conference of Eastern Asia Society for Transportation Studies,* Jeju, Korea, June 20-23.

(2) 國際期刊論文 1 篇：

Chiou, Y.C., Lan, W.L. and Chen, W.B. (2010) "Contributory factors to crash severity in Taiwan freeways: Genetic mining rule approach," *Journal of Eastern Asia Society for Transportation Studies*, Vol.8, pp.1837-1849.

(3) 投稿期刊論文 2 篇：

Chiou, Y.C., Lan, W.L. and Chen, W.B. (2011) "Identification and estimation the risky conditions of crash severity: An integrated genetic rule mining and ordered probit model," submitted to *Accident Analysis and Prevention.*

Chiou, Y.C., Hwang, C.C. Chang, C.C. and Fu, C. (2011) "A bivariate generalized ordered probit approach for two-party crash severity modeling," Invited submission to *Accident Analysis and Prevention* (Special Issue of RSS2011)

(4) 博士論文寫作 1 名（進行中）：

陳文斌，Identifying risk conditions to crash severity by a two-stage model combining genetic rule mining and ordered Probit，交通大學交通運輸研究所，博士論文（進行中），民國 100 年。

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

本計畫為三年期計畫，具體完成之研究成果如下：

1.建立基因規劃探勘模式

本計畫乃於第一年期提出基因規則探勘模式（Genetic rule mining, GRM），可由探勘所得之規則的前半部，判斷何謂危險情況，進而加以避免。惟本研究所提出之 GRM 必須先固定規則數量，再同時進行最佳規則組合之尋優。因此，具有染色體長度過長，尋優效果不佳，以及探勘過多衝突或重覆規則的傾向，進而導致規則難以詮釋，無法提出具體之安全改善策略。

2.提出改良式逐步基因規劃探勘模式

本計畫第二年期乃提出改良式的基因規則探勘模式（Genetic rule mining, GRM），稱為逐步基因規則探勘模式（Stepwise GRM, SGRM）。SGRM 一次僅挑選使事故嚴重度預測率精確率最高的一條規則，再以此規則為基礎，進行下一條規則之選取，直到精確率無法再改善為止。如此，即可避免選擇規則過多，且相互重覆或矛盾的問題。此外，由於不同類型事故之影響因素與危險情況不一定相同，因此，有必要加以區隔分析。本年度以先以總計 5563 件單車事故（single vehicle accident）為分析基礎。結果顯示，本模式共選擇了 38 條規則，其訓練準確度達 75.1%，而驗證準確度則達 73.8%均遠高於決策樹之預測結果。而影響事故嚴重度之危險情境也加以確認，並研提改善策略。

3.進行模式比較與分析

本計畫第三年期進一步探討及比較本研究所提出之 SGRM、決策樹（decision tree, DT）、粗略集合（rough set, RS）及次序普羅比（ordered probit model, OP）等四種模式在分析不同事故嚴重度之選擇規則與重要解釋變數。結果顯示，本研究所提出之 SGMR 模式不僅可達到最佳之預測準確度外，也可篩選出少數關鍵推理規則及危險情境，俾利安全改善策略之研提。

4.提出兩階段整合模式

本研究進一步將 SGRM 所挑選出的 38 條推理規則的前半部設定為危險情境（risk condition），以虛擬變數表之，後半部則為事故嚴重度，結合次序普羅比進行危險情境之推估與檢定，以了解各種危險情境對事故嚴重度之影響程度。結果顯示，本研究所提出之整合方法（SGMR＋OP）其模式配適度，遠比將所有原始變數作為解釋變數所建構之 OP 模式為高，更可有效辨識、檢定及推估各種危險情境，有效克服以往統計迴歸方法僅能探討單一變數對事故嚴重度的缺點，更符合事故嚴重度係由多個肇因所導致之先驗知識。

四、參考文獻

1. Abdel-Aty, M. and Pande, A., 2007. Crash data analysis: Collective vs. individual crash level approach. *Journal of Safety Research* 38, 581-587.
2. Abdel-Aty, M. and Radwan, A.E., 2000. Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention* 32, 633-542.
3. Al-Ghamdi, A.S., 2002. Pedestrian-vehicle crashes and analytical techniques for stratified contingency tables. *Accident Analysis and Prevention* 34, 205-214.
4. 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.
5. 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.
6. 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.
7. Chen, T.C. and Hsu, T.C., 2006. A GAs-based approach for mining breast cancer pattern. *Expert Systems with Applications* 30, 674-681.
8. 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.
9. 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.
10. Dehuri, S. and Mall, R., 2006. Prediction and comprehensible rule discovery using a multi-objective genetic algorithm. *Knowledge-Based System* 19, 413-421.
11. 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.
12. 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.
13. Freitas, A.A., 1999. On rule interestingness measures. Knowledge-Based Systems 12, 309-315.
14. Ghosh, A. and Nath, B., 2004. Multi-objective rule mining using genetic algorithms. *Information Sciences* 163, 123-133.
15. Goldberg, D.E., 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, New York.
16. Greibe, P., 2003. Accident prediction models for urban roads. *Accident Analysis and Prevention* 35, 273-285.
17. 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.
18. 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.
19. Kim, M.J. and Han, I., 2003. The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. *Expert Systems with Applications* 25, 637-646.
20. Michalewicz, Z., 1992. *Genetic Algorithms + Data Structures = Evolution Programs*, Springer, Berlin.
21. 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.
22. Poch, M. and Mannering, F., 1996. Negative binomial analysis of intersection. *Journal of Transportation Engineering* 12, 105-113.
23. Shanker V., Mannering, F. and Barfield, W., 1996. Statistical analysis of accident severity on

rural freeways, *Accident Analysis and Prevention* 28, 391-401.
24. Shin, K.S. and Lee, Y.J., 2002. A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications* 23, 321-328.
25. 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.
26. Tay, R. and Rifaat, S.M., 2007. Factors contributing to the severity of intersection crashes. *Journal of Advanced Transportation* 41, 245-265.
27. 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.
28. Chiou, Y.C., Lan, W.L. and Chen, W.B. (2009) "Contributory factors to crash severity in Taiwan freeways: Genetic mining rule approach," *presented at the 8th International Conference of Eastern Asia Society for Transportation Studies,* Surabaya, Indonesia, Nov. 16-18.
29. Chiou, Y.C., Lan, W.L. and Chen, W.B. (2010) "Contributory factors to crash severity in Taiwan freeways: Genetic mining rule approach," *Journal of Eastern Asia Society for Transportation Studies*, Vol.8, pp.1837-1849.
30. Chiou, Y.C., Lan, W.L. and Chen, W.B. (2010) "Identification of risky conditions contributing to crash severity with genetic mining rules," *presented at the 15th Conference of Hong Kong Society for Transportation Studies,* Hong Kong, China, Dec. 11-14.
31. Chiou, Y.C., Hwang, C.C. Chang, C.C. and Fu, C. (2011) "A bivariate generalized ordered probit approach for two-party crash severity modeling," *presented at the 3rd International Conference on Road Safety and Simulation*, September 14-16, 2011, Indianapolis, USA. (Invited submission to *Accident Analysis and Prevention*)
32. Chiou, Y.C., Lan, W.L. and Chen, W.B. (2011) "Exploring key discriminative and risk factors affecting one-vehicle crash severities in Taiwan freeways," *presented at the 9th International Conference of Eastern Asia Society for Transportation Studies,* Jeju, Korea, June 20-23.
33. Chiou, Y.C., Lan, W.L. and Chen, W.B. (2011) "Identification and estimation the risky conditions of crash severity: An integrated genetic rule mining and ordered probit model," submitted to *Accident Analysis and Prevention.*
34. Chiou, Y.C., Hwang, C.C. Chang, C.C. and Fu, C. (2011) "A bivariate generalized ordered probit approach for two-party crash severity modeling," Invited submission to *Accident Analysis and Prevention* (Special Issue of RSS2011)
35. 陳文斌，Identifying risk conditions to crash severity by a two-stage model combining genetic rule mining and ordered Probit，交通大學交通運輸研究所，博士論文（進行中），民國100年。