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 (I/III)

Developing Genetic and Ant-based Rule Mining Models- Case Studies on Accident Analysis and Appraisal (I/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 副教授

計畫參與人員：陳彥蘅、傅強　　交通大學交研所博士班研究生

林柏辰、謝志偉　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 98 年 5 月 25 日

一、摘要

1.1 中文摘要

本計畫第一年期之研究內容旨在利用規則探勘模式（genetic rule mining, GRM），建立一套事故鑑定專家系統，除可供事故責任鑑定之精確預測外，並提供學習所得規則，供進一步詮譯、後最佳化及訓練之用。本研究共提出三種 GRM 模式，並利用民國 89～91 年之事故鑑定案例，共計 537 件，1,074 位當事人作為研究對象，透過卡方檢定共評選出 22 個關鍵鑑定變數，以作為潛在之狀態變數，控制變數則設定為鑑定責任（分別全因、主因、次因、同為及無因等五個等級）。所有變數均設定為類別變數（categorical variables），俾利變數數值之選擇。至於 GAs 染色體之編碼，則以每一條染色體代表一條邏輯規則，每一族群代表所組成之邏輯規則組成。各染色體之適合度值（fitness value）則以三個目標表之，分別為：包含事故件數最多、與其他染色體重覆最少、與其他染色體矛盾度最低。並與類神經網路及判別分析方法進行績效評比。結果顯示，本模式之判中率明顯高於類神經網路及判別分析模式，且所選擇之邏輯規則也可據以了解鑑定委員之推理邏輯。

另外，本年期也進一步將所建構之 GRM 模式應用於高速公路事故嚴重度之分析。為避免三級嚴重度之事故件數分佈不均所導致之過度挖掘問題，本研究乃以隨機方式選擇相同數量之事故件數進行模式訓練與驗證，並將預測結果與決策樹比較，以驗證本模式之績效。結果顯示，本模式共計選擇了 19 條規則，其訓練準確度達 78.50%，而驗證準確度則達 74.16%均遠高於決策樹之預測結果。而障礙物及鋪面狀況是兩個最重要的影響事故嚴重度之成因。

關鍵字：事故鑑定、規則探勘、遺傳演算法、事故嚴重度

1.2 Abstract

The first research year of this project aims to propose an accident appraisal expert system that can not only accurately predict the liability degrees of involved parties but also demonstrate comprehensive appraisal rules for further investigation, post-adjustment, and training of junior accident appraisal committee members. Three genetic rule mining (GRM) models, based on two schemes of the Michigan approach (GRM1 and GRM2) and on the Pittsburgh approach (GRM3) are respectively proposed by discovering knowledge from historical appraisal cases. A total of 537 Taiwanese two-car crash accident cases (1,074 parties) are randomly and equally divided into three subsets to train and validate the proposed models. The GRM1 model performs the best in both training and validation with correctness rates of 78.85% and 70.21%, respectively. We further compare the proposed GRM1 model with artificial neural network models (ANN) and discrimination analysis (DA) model proposed by Chiou (2006). The GRM1 model can achieve the same accuracy as the ANN models and provide more information than the ANN models by delivering a comprehensive combination of rules. It can be used to enhance the quality and efficiency of accident appraisal.

In addition, the proposed GRM models have also been applied to freeway accident analysis to discover the key rules that determine the most contributing factors to crash severity. To avoid over-mining caused by unevenly distributed data across different types of accidents, identical numbers of A1-type, A2-type, and A3-type crash cases drawn from 2003-2007 Taiwan freeway accident investigation reports are used for the analysis. A total of 19 rules have been mined which can achieve overall correct rates of 78.50% in training and of 74.16% in validation, respectively, much higher than those yield by the decision tree model. Obstacle and surface condition have been found as the two most contributory factors to crash severity in this study.

Keywords: Accident appraisal; Rule mining; Genetic algorithms; Crash severity.

二、主要研究成果

本年期主要研究成果包括兩大部分：事故鑑定及事故分析。分述如下：

2.1 事故鑑定

2.1.1 Background

Almost all countries have official institutes or ad hoc committees responsible for investigating the road traffic accident liabilities. In Taiwan, two sorts of such ad hoc committees have long been in operation: the local appraisal committee (LAC) and the re-appraisal committee (RAC). The LAC is responsible for investigating the liability of disputable cases of road accidents taking place within a jurisdiction area; whereas the RAC is authorized to re-investigate the cases that the LAC's judgments are not agreeable to any of the involved parties within a region covering one or several LAC territories. Nowadays, there are 14 LACs and 5 RACs are in operation in Taiwan but several defects have been identified (Chiou, 2006). The most serious problem is the insufficient number of experienced experts because carrying out the accident appraisals requires professional knowledge. There has been lack of mechanism to pass the cumulated experiences of senior members over to the new members when the terms for senior members are expired or when the committees are reshuffled periodically. As a consequence, it is not unusual that the appraisal or re-appraisal outcomes for very similar cases judged by different LACs or RACs can be quite diverse or even contradictory. It is imperatively important but challenging to develop effective expert systems that can help enhance the consistency of the accident appraisals.

Chiou (2006) developed an artificial neural network (ANN)-based accident appraisal expert system wherein two models (party-based and case-based) are proposed and compared. It is found that the party-based model has reached correctness rates of 82.43% (training) and 68.75% (validation), while the case-based model has achieved 85.72% (training) and 77.91% (validation). With such satisfactory correctness rates, the models are in effect of practical helpfulness. However, after implementing the proposed ANN models to some RAC members, two major concerns have been pointed out. First, the committee members express their hesitations to use the black-box characterized ANN models because they fail to clearly get insights into the ANN inference mechanism. Second, they question the capability of ANN models in training the junior members because the models only predict the liability with lack of explanations. It would be of great usefulness if one could develop an expert system that can not only achieve higher correctness rates but also convey the knowledge extracted from historical cases to any committee members, i.e., mining for knowledge from available historical databases and toward decision support of accident appraisals.

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). Rule mining problems can be roughly divided into two categories: crisp rule mining and fuzzy rule mining, depending upon the fuzziness of the variables in rules. 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. In combining with GAs, two categories of rule mining algorithms can be found in literature: genetic mining rule (GMR) (*e.g.* Freitas, 1999; Ghosh and Nath, 2004; Dehuri and Mall, 2006) and genetic fuzzy logic controller (GFLC) (*e.g.* Herrera, *et al.*, 1995, 1998; Lekova, 1998; Chiou and Lan, 2005). The performances of these rule mining algorithms have been proven and applied in many fields. Since the data in the accident appraisal cases are crisp and categorical in nature, GMR is more suitable for accident appraisal purposes than GFLC. Thus, this paper will develop GMR models that can determine the optimal combination of appraisal rules to achieve the following goals: (1) to accurately predict the liability degree of the involved parties (as an expert system); (2) to train the

junior LAC or RAC committee members (as a training tool); and (3) to provide the possibility of post-adjustment (fine-tune) of the rules mined. Previous relevant studies have seldom considered the problems of conflicts and redundancy among the mined rules, our proposed GMR models will account for conflicts and redundancy in addition to conventional objectives: coverage ratio and predictive accuracy.

2.1.2 Data

In Taiwan, the ad hoc committee members at the LAC or RAC levels assess accidents based on the investigation reports prepared by the police at accident sites. These reports are finalized with some tables, figures, photos, and scripts. The information of the reports are digitized and divided into five categories with 39 variables, as shown in Table 1. It includes (1) the background of the accident (*e.g.*, date, time, location, type of road, daylight or darkness, weather condition, speed limit); (2) demographics of the drivers and characteristics of the vehicles (gender, age, education, type of vehicle, length of vehicle); (3) violations (licensing, speeding, invasion, alcohol use); (4) behaviors of drivers (*e.g.*, direction, movement, foresight of the accident); (5) evidence (braking line of left and/or right wheel, crash spot, self-reported speed, driver injury, passenger injury, driver death, passenger death). In each accident case, the committee members summarized their appraisal report based on the police investigation report. If a consensus had been reached in the committee, the appraisal report was finalized with a clear statement of the accident liabilities for all parties involved, with a full explanation of the reasons. The liabilities (y) of all parties are categorized into five degrees: full responsibility (y=5), *i.e.*, the one who had to take complete responsibility for causing the accident, major responsibility (y=4), equal responsibility (y=3), minor responsibility (y=2), and no responsibility (y=1).

In order to generate appraisal rules, the same dataset of historical accident appraisal cases studied by Chiou (2006) is adopted, which is composed of 537 cases, involving 1,074 parties, of two-car crash cases with sufficient information indicated in Table 1 and with consistent appraisal results between the LAC and the RAC. These cases are selected from an original dataset of 5,641 historical appraisal cases during the period 2000-2002 from the Taiwan Provincial RAC. For training and validation purposes, these 537 cases are randomly and equally divided into three subsets, each of which consists of 179 cases (358 parties).

Table 1
Accident digitalized data summarized from police investigation report

Category	Information	Variable	Coding	Descriptions
Background	Date	x_1	Character	month/date/year
	Time	x_2	Character	hour/minutes
	Type of road	x_3	Categorical	1, national freeway; 2, provincial highway; 3, county highway; 4, rural highway; 5, street
	Location	x_4	Categorical	1, straight road; 2, curved road; 3, signalized intersection; 4, flashlight intersection; 5, not signalized intersection
	Major or minor street	x_5	Categorical	1, major street; 2, minor street; 3, not clear
	Lane located	x_6	Categorical	1, inner lane; 2, outer lane; 3, middle lane; 4, slow lane; 5, one way street
	Day or night	x_7	Categorical	1, day; 2, night with illumination; 3, night without illumination
	Weather condition	x_8	Categorical	0, clear; 1, rainy or cloudy,
	Flash signal	x_9	Categorical	1, flash red; 2, flash yellow; 3, no flash signal
	Speed limits	x_{10}	Continuous	km/hr
Demographics	Gender of driver	x_{11}	Categorical	1, male; 2, female
	Age of driver	x_{12}	Integer	years
	Education	x_{13}	Categorical	1, university; 2, college; 3, high school; 4, high vocational school; 5, junior high school; 6, elementary school; 7: kindergarten
	Type of vehicle	x_{14}	Categorical	1, passenger car; 2, business car; 3, light truck; 4, truck; 5, bus
	Length of vehicle	x_{15}	Continuous	Meter
Violations	Licensing	x_{16}	Categorical	1, yes; 2, no (above licensing age); 3, no (below licensing age)
	Speeding	x_{17}	Categorical	1, seriously speeding (over 20km/hr); 2, speeding; 3, no
	Invasion	x_{18}	Categorical	1, invasion to opposing direction; 2, moving backward; 3, no

				violation; 4, not clear; 5, not follow the signal or markings
	Alcoholic use	x_{19}	Categorical	1, yes (>0.55mg/l); 2, yes (0.25mg/l~ 0.55mg/l); 3, yes (< 0.25mg/l); 4, no
Behaviors	Movement	x_{20}	Categorical	1, forward; 2, right turn; 3, left turn; 4, u turn; 6, stop; 7, backward
	Direction	x_{21}	Categorical	1, east to west; 2, west to east; 3, south to north; 4, north to south
	Lane change	x_{22}	Categorical	0, no; 1, yes; 2, overtaking
	Foresight of the accident	x_{23}	Categorical	0, no; 1, yes; 2, not clear
	Foresight distance	x_{24}	Continuous	meter
	Reactions	x_{25}	Categorical	0, no; 1, flash; 2, flash right; 3, flash left; 4, lane change; 5, reverse; 6, detour; 7, horn; 8, flash light; 9, decelerate; 10, stop; 11, pass; 12, not clear; 13, escape
	Braking	x_{26}	Categorical	0, no; 1, brake before crash; 2, brake after crash
Evidences	Braking line of left wheel	x_{27}	Continuous	meter
	Braking line of right wheel	x_{28}	Continuous	meter
	Related direction	x_{29}	Categorical	1, opposing direction; 2, same direction; 3, left adjacent direction; 4, right adjacent direction
	Crash spot	x_{30}	Categorical	0, no damage; 1, right front; 2, right-hand side; 3,left rear; 4, rear; 5, left rear; 6, left-hand side; 7, left front; 8, front
	Self-reported speed	x_{31}	Continuous	km/hr
	Relative position	x_{32}	Categorical	1, in the front; 2, in the rear; 3, in the left; 4, in the right; 5, start from roadside; 6, opposing direction
	Crossing the middle of intersection	x_{33}	Categorical	1, no; 2, yes; 3, not at an intersection
	Number of lanes after turn	x_{34}	Categorical	1, one; 2, two; 3, more than two
	Lane after turn	x_{35}	Categorical	1, inner lane; 2, outer lane; 3, middle lane; 4, slow lane; 5, one-way street
	Driver injury	x_{36}	Categorical	0, no; 1, yes
	Passenger injury	x_{37}	Integer	Persons
	Driver death	x_{38}	Categorical	0, no; 1, yes
	Passenger death	x_{39}	Integer	Persons

Source: Chiou (2006)

The cases appealed to Taiwan Provincial RAC for reappraisal have been previously assessed by one of the 12 LACs, which consist of completely different committee members. To examine the discrepancy of appraisal results concluded by different LACs, a variable (x_{40}), with values of 1-12 representing the different LACs, is added. Furthermore, right-of-way, the priority in using the road, is a critical factor in assessing the liability of involved parties. However, such a judgment is too professional for the police officers at site to be concluded in the investigation report. Thus, Chiou (2006) propose a total of 38 decision trees to determine the right-of-way of involved party according to 12 out of the 39 variables in Table 1, including direction, movement, lane located, lane after turn, number of lanes after turn, relative positions, crossing the middle of the intersection, and violations, etc. The right-of-way is a dummy variable (x_{41}) with values of 1 and 2. $x_{41} = 1$ indicates that party is assessed to own the right-of-way; $x_{41} = 2$ otherwise.

In addition, to simultaneously describe the situation of the two drivers (vehicles) involved in an accident, a superscript is further added to each variable for each of driver in developing case-based models. That is, x_i^1 and x_i^2 represent the i^{th} variable corresponding to driver 1 (vehicle 1) and driver 2 (vehicle 2), respectively. Likewise, y^1 and y^2 represent the appraisal results (responsibility degrees) of driver 1 (vehicle 1) and driver 2 (vehicle 2), respectively.

These 41 variables are not all closely related to liability assessment. The correlated relationships between exploratory variables and assessed liability should be first examined to remove unrelated variables and reduce the number of potential rules. Since most of the explanatory variables and appraisal results are categorical, the table of contingency technique is adopted to investigate the significant relationships among them. The results are presented in Table 2, in which 12 variables are selected at the 0.05 significance level. To facilitate the rule mining process, all categorical variables are re-coded with values started from 1, instead of 0 and all continuous variables (only one in this case) are categorized into several classes.

Table 2
Variables selected by table of contingency

Variable	Original notation	New notation	Value
Type of road	x_3	z_1	1, national freeway; 2, provincial highway; 3, county highway; 4, rural highway; 5, street.
Location	x_4	z_2	1, straight road; 2, curved road; 3, signalized intersection; 4, flashlight intersection; 5, unsignalized intersection.
Type of vehicle	x_{14}	z_3	1, passenger car; 2, business car; 3, light truck; 4, truck; 5, bus.
Speeding	x_{17}	z_4	1, seriously speeding (over 20 km/hr); 2, speeding; 3, no.
Alcoholic use	x_{19}	z_5	1, yes (>0.55mg/l); 2, yes (0.25mg/l~ 0.55mg/l); 3, yes (< 0.25mg/l); 4, no.
Direction	x_{21}	z_6	1, east to west; 2, west to east; 3, south to north; 4, north to south.
Foresight of the accident	x_{23}	z_7	1, none; 2, yes; 3, not clear.
Crash spot	x_{30}	z_8	1, no damage; 2, right front; 3, right-hand side; 4,left rear; 5, rear; 6, left rear; 7, left-hand side; 8, left front; 9, front.
Self-reported speed	x_{31}	z_9	1, < 31 km/hr; 2, 31-40 km/hr; 3, 41-50 km/hr; 4, 51-60 km/hr; 5, 61-70km/hr; 6, > 70km/hr; 7, not clear.
Driver death	x_{38}	z_{10}	1, no; 2, yes.
Area of LAC	x_{40}	z_{11}	1-12 corresponding areas of LAC.
Right-of-way	x_{41}	z_{12}	1, yes; 2, no.

Note: The variable, self-reported speed, which is originally coded as continuous values are then categorized into seven classes. Three variables, foresight of the accident, crash spot, driver death, which are originally coded with values starting from 0, are re-coded as values starting from 1.

2.1.3 Methodologies

The Pittsburgh approach and the Michigan approach are the two main approaches to encoding rules. The former is a natural way to represent an entire rule set as a chromosome, maintain a population of candidate rule sets. Historically, this was the approach taken by DeJong and his students while at the University of Pittsburgh, which gave rise to the phrase “the Pittsburgh approach” (Smith, 1983; DeJong, 1988). The fitness value of a chromosome in such an approach can be directly represented as the performance index of the entire rule set, but the number of rules mined and the length of a rule of this approach are strictly restrained by the length of chromosome. In contrast, the latter is a model of cognition in which the members of the population are individual rules and a rule set is represented by the entire survived population. This approach was originally taken by Holland and his students while at the University of Michigan, therefore, the approach is named as the Michigan approach (Holland and Remitman, 1978; Booker *et al.*, 1989). Compared with the Pittsburgh approach, the Michigan approach can accommodate larger number of rules and literally lengthy rules, but the fitness value of a chromosome, *i.e.* a rule, is much more difficult to define. Because an expert system constituted of top-performing rules in a survived population may not necessarily perform well on the whole, if these rules are conflicting or redundant to each other, thus it is challenging to design an appropriate fitness function to measure the performance of a chromosome (a rule) such that the entire mined (survived) rules can perform most excellent on the whole. The proposed encoding methods and fitness functions of these two approaches are respectively described below.

2.1.3.1 Encoding methods

(1) Michigan approach

In the Michigan approach, each chromosome represents a candidate If-then rule. The conditions associated in the "if part" are termed as antecedent and those in the "then part" are named consequent. Besides, the antecedent part consists of at least one to at most twelve variables selected from Table 2 and the consequent part is composed by, of course, only one variable: liability degree of involved party. In general a rule is a knowledge representation of the form If A then C, where A is a set of parties satisfying the conjunction of predicting attribute values and C is a set of parties with the same predicted class. Thus, a typical rule i can be of the form:

Rule i: If $z_1=a_{i1}$ and $z_2=a_{i2}$ …and $z_j=a_{ij}$ … and $z_{12}=a_{i12}$ Then $y=g_i$

or of 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 5 to represent five degrees of liability. A_i and C_i, again, are the sets of parties satisfying the antecedent part and consequent part of rule i, respectively.

By encoding a rule as a chromosome, each gene is used to represent a corresponding variable. The length of a chromosome is set as 13 to represent twelve state variables in the antecedent part and one control variable in the consequent part. Each gene takes one of the categorical values of the corresponding variable. It is worthy of noting that the ranges of gene value may differ from each other since the range of corresponding variable is different. In addition, a gene in a rule antecedent is allowed to take a value of 0 to represent that its corresponding variable is not considered by the rule. If all genes in the antecedent part simultaneously take a value of 0 or the gene in the consequent part takes 0, then the whole rule is not included in the rule set.

Based on such an encoding method, a rule of "If speeding="no" and alcoholic use="no" and right-of- way="yes", then liability="no responsibility" can be encoded as 0003400000011 as depicted in Fig. 1. The total number of potential rules equals to $6\times6\times6\times4\times5\times5\times4\times10\times8\times3\times13\times3\times5$ =4,043,520,000, making the number of potential rule combinations reaches $2^{4,043,520,000}$. Obviously, it is barely possible to compare all rule combinations by a total enumeration manner.

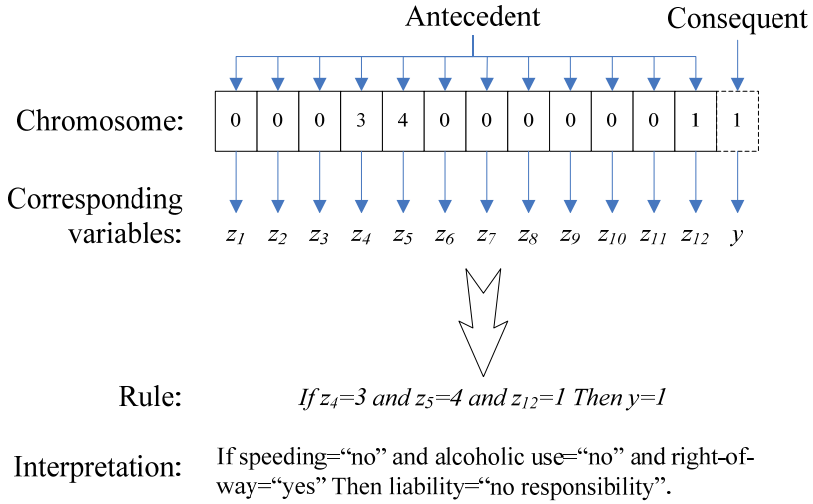
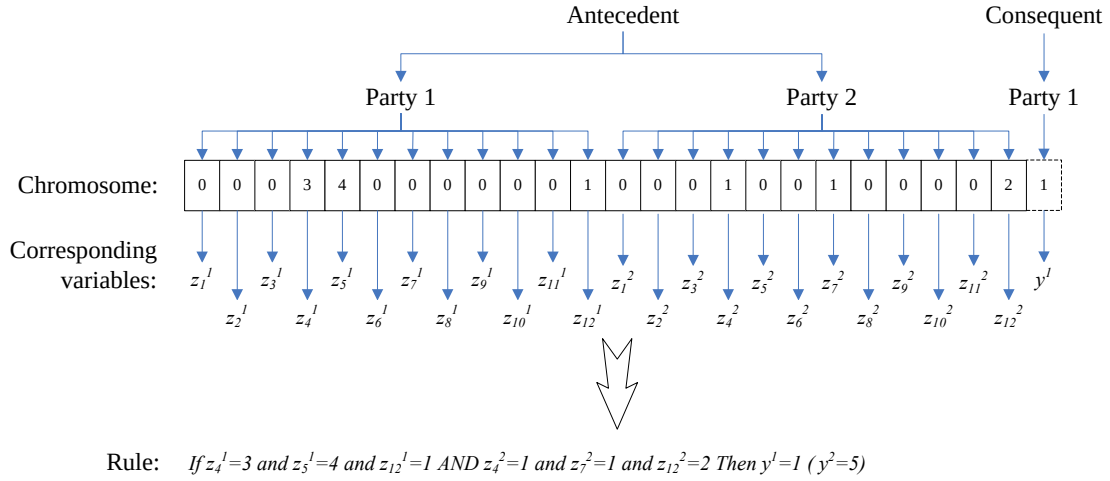


Fig. 1. Encoding methods of party-based Michigan model (GRM1)

In the context of two-car crash accidents, it is interesting to compare the performance of the models which simultaneously consider the state variables of both parties involved or only consider that of one party alone. Therefore, two GRM models, namely GRM1 and GRM2, are proposed. The GRM1 is a party-based Michigan approach model which considers the state variables of one party alone to conclude the liability degree for the party as shown in Fig.1, while GRM2 is a case-based Michigan approach model which considers both parties involved simultaneously to conclude the liability degree of one party. Since the sum of liability degrees of two parties, party 1 and party 2,

involved in an accident are always equal to a constant, *i.e.* 6. For instance, of party 1 is judged as no responsibility for the accident, then party 2 must have to take full responsibility. The chromosome length of the GRM2 is 25 as depicted in Fig. 2.



Rule: *If z_4^1=3 and z_5^1=4 and z_{12}^1=1 AND z_4^2=1 and z_7^2=1 and z_{12}^2=2 Then y^1=1 (y^2=5)*

Interpretation: If speeding of party 1="no" and alcoholic use of party 1="no" and right-of-way of party 1="yes" AND speeding of party 2="seriously speeding" and foresight of the accident of party 2="yes" and right-of-way of party 1="no" Then liability of party 1="no responsibility" (implying that liability of party 2="major responsibility")

Fig. 2. Encoding methods of case-based Michigan model (GRM2)

(2) Pittsburgh approach

In contrast to the Michigan approach, the Pittsburgh approach simultaneously encodes all mined rules, a rule set, as a chromosome by simply extending the length of a chromosome to a multiple length of the Michigan approach's chromosome of the number of rules. Therefore, the quality of the chromosome, *i.e.* fitness function, can straightforwardly represent the correctness rate of the rule set. Taking a preset number of rules (K)=20 for instance, the party-based encoding method (hereinafter, named as GRM3) can be depicted as Fig. 3. The length of chromosome equals to $13 \times 20 = 260$. Obviously, the case-based Pittsburgh approach model will result in a very lengthy chromosome, thus it is not considered in this paper. Since the situations that all antecedent genes or the consequent gene in a rule take a value of 0, then it indicates that the rule will not be selected. In other words, the number of mined rules must be less than or equal to the preset number of rules (K).

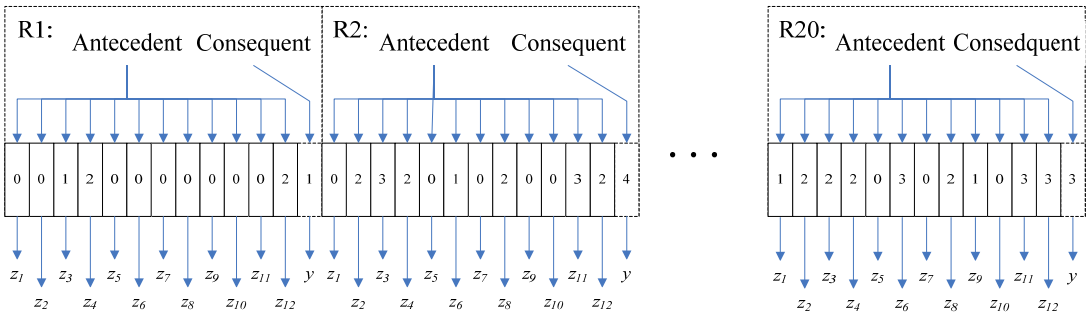


Fig. 3. Encoding methods of party-based Pittsburgh model (GRM3)

2.1.3.2 Performance indices

For the Michigan approach, the role of the fitness function is to evaluate the quality of the rule numerically. In doing so, three common factors to be taken into account are the coverage, the completeness and the confidence factor of the rule, respectively defined as follows. The coverage of the rule, *i.e.* the cases satisfied by the rule antecedent, is given 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 the rule, *i.e.* the predictive accuracy, is given by $|A \cap C|/|A|$ (Freitas, 1999). Based upon these three factors, this paper uses the performance indices including predictive accuracy, comprehensibility and interestingness narrated as follows.

(1) Predictive accuracy

Two ways to measure the predictive accuracy, also called confidence, are found in literature (Ghosh and Nath, 2004, Dehuri and Mall, 2006, Dehuri *et al*., 2008):

$$P(\Re) = |A \cap C|/|A|) \tag{1}$$

$$P(\Re) = \frac{|A \cap C| \times |\sim A \cap \sim C|}{(|A \cap C| + |A \cap \sim C|) \times (|A \cap C| + |\sim A \cap C|)} \tag{2}$$

where $P(\Re)$ is the predictive accuracy of rule $\Re$. A is a set of cases or parties satisfying the rule antecedent. C is a set of cases or parties satisfying the rule consequent. $\cap$ is the intersection operator. $\sim A$ and $\sim C$ are complement sets of A and C, respectively. $|A|$ is the cardinality of set A.

(2) Comprehensibility

Comprehensibility is used to measure the family size of the rule. The smaller the rule, the more comprehensible (specific) it is. There are several ways to measure comprehensibility (Fedelis, 2000, Ghosh and Nath, 2004, Dehuri and Mall, 2006, Dehuri *et al.*, 2008) for example:

$$C(\Re) = 1 - N_c(\Re)/M_c \tag{3}$$

$$C(\Re) = M_c - N_c(\Re) \tag{4}$$

$$C(\Re) = N - |A| \tag{5}$$

$$C(\Re) = \log(1+k)/\log(1+m+k) \tag{6}$$

where $C(\Re)$ is the comprehensibility of rule $\Re$. $N_c(\Re)$ is the number of conditions in the rule $\Re$. M_c is the number of at most conditions a rule can have. N is the number of total cases. m and k are the number of attributes involved in the antecedent part and consequent part, respectively.

(3) Interestingness

The third criterion of the rules, called interestingness, is used to measure how surprising, useful or novel the rule is. Two simpler expressions of interestingness can be found in literature (Piatetsky-Shapiro, 1991, Ghosh and Nath, 2004) as follows:

$$I(\Re) = |A \cap C| - |A||C|/N \tag{7}$$

$$I(\Re) = \frac{|A \cap C|}{|A|} \times \frac{|A \cap C|}{|C|} - (1 - \frac{|A \cap C|}{N}) \tag{8}$$

where $I(\Re)$ is the interestingness of rule $\Re$. N is the total number of cases or parties.

2.1.3.3 Proposed fitness functions

(1) GRM1 and GRM2

Many studies consider rule mining as a multi-objective problem and employ evolutionary algorithms to mine the Pareto-optimal rules with respect to abovementioned three indices. However, in mining classification rules instead of association rules, a combination of Pareto-optimal rules

might not necessarily perform best on the whole. Besides, the conflict and redundancy among mined rules are seldom considered in the literature. High redundancy and conflict among rules will cause too many similar or conflicting rules being mined.

Based on this, this paper designs a two-stage process to select chromosomes (rules). In the first stage, a fitness function, a performance index with a combination of coverage ratio, predictive accuracy, and predictive error, is defined and used to rank the chromosomes, which can be expressed as:

$$f_i = \frac{|A_i|}{N} \times \left(\frac{|A_i \cap C_i|}{|A_i|} - \frac{|A_i \cap \sim C_i|}{|A_i|}\right) \tag{10}$$

where, f_i is the performance index of the i^{th} chromosome. The first term of right hand side is coverage ratio. The second term is the predictive accuracy rate. The third term is predictive error rate. The equation can be then simplified as:

$$f_i = \frac{|A_i \cap C_i|}{N} - \frac{|A_i \cap \sim C_i|}{N} \tag{11}$$

In the second stage, to avoid selecting similar or conflicting rules, two indices, redundancy index and conflicting index, are respectively defined and then both used to filter high redundant and conflicting chromosomes (rules). The redundancy index of rule i is defined as

$$l_i = \max_{j=1}^{i-1}\left\{\frac{|A_i \cap A_j|}{|A_i \cup A_j|}\right\} \tag{12}$$

where l_i is the redundancy index of the i^{th} chromosome which has already been ranked in a descending order. The conflicting index of rule i is defined as

$$t_i = \max_{j=1}^{i-1}\left\{\frac{|A_i \cap A_j \text{ if } g_i \neq g_j|}{|A_i \cup A_j|}\right\} \tag{13}$$

where t_i is the redundancy index of the i^{th} chromosome which has been ranked in a descending order. $|A_i \cap A_j \text{ if } g_i \neq g_j|$ is the number of cases or parties satisfying both antecedent of rules i and j which have different predicted classes.

To facilitate the filtering process according to l_i and t_i, two thresholds, L and T, for redundancy and conflict have to be given. Thus, if a chromosome reaches either $l_i \geq L$ or $t_i \geq T$, then it will be replaced by another randomly re-generated chromosome (newly re-born chromosome). These two given thresholds will then determine how many chromosomes will survive in each of generation. The higher the values of the thresholds are, the more chromosomes will be replaced. Of course, if too many chromosomes are replaced by randomly generated chromosomes, the competitive genetics can not be successfully preserved and a randomly searching mechanism will be resulted. Thus, the values of these two thresholds should be carefully examined. It is worthy of noting that the redundancy and conflicting indices are computed through a mutual comparison matter by only comparing with the chromosomes ranked in front.

The evolutional process of GRM1 and GRM2 can be depicted in Fig. 4. As shown from the figure, after selection, crossover, and mutation, the survived chromosomes are first ranked according to their fitness value, and then filtered by their redundancy and conflicting indices. In the process all highly redundant and conflicting chromosomes will be replaced by randomly generated ones until the stopping condition (a given number of generations in this paper) reaches. However, in the final generation, all highly redundant and conflicting chromosomes will be deleted without replacement. The finally survived chromosomes are the optimal combination of rules.

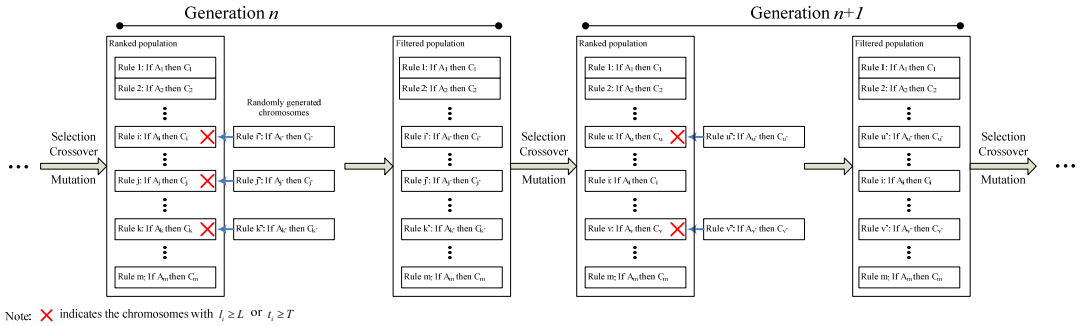


Fig. 4. Evolutional process of the GRM1 and GRM2 models

Even if the chromosomes have been filtered by redundancy and conflicting indices, it can not be avoided that two or more rules with different predicted classes might still be simultaneously fired by a sample. To synthesize the predicted class of more then one rules fired, we take an average value of predicted classes of all fired rules and round it to a nearest integer, which can be expressed as:

$$sg = Int(\frac{1}{|F|}\sum_{j\in F} g_j) \tag{14}$$

where, G is the predicted class of the algorithm. $Int(\cdot)$ is a rounding operator, which rounds value in parenthesis to a nearest integer. F is a set of sequence numbers of fired rules. As such, the correctness rate of the model can be then computed as the number of correctly predicted cases or parties (that is, the predicted class equal to the target class) divided by total number of cases or parties.

(2) GRM3

Since a chromosome of the GRM3 model (the party-based Pittsburgh approach) represents a combination of rules, the fitness function of a chromosome can be directly expressed as correctness rate. Correctness rate is defined as the number of correctly predicted parties divided by the total number of parties in training or validation. In the case of more than one rule fired, Eq.(14) is also used to synthesize the predicted classes of all fired rules. The reasons for not taking redundancy and conflicting indices into account are that the GRM3 aims to maximize the correctness rate under a preset number of rules (K), at optimality, the redundancy and conflicts among rules should be largely avoided.

2.1.3.4 Genetic operators

Because the genes in our GRM models are not encoded binary, simple genetic algorithms (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 will be generated:

$$G_1^{t+1} = aG_w^t + (1\text{-}a)G_v^t \tag{15}$$

$$G_2^{t+1} = aG_v^t + (1\text{-}a)G_w^t \tag{16}$$

$$G_3^{t+1} \text{ with } g_{3k}^{t+1} = min\{g_{wk}^t, g_{vk}^t\} \tag{17}$$

$$G_4^{t+1} \text{ with } g_{4k}^{t+1} = max\{g_{wk}^t, g_{vk}^t\} \tag{18}$$

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

(2) Non-uniform mutation

Let $G_t = \{ g_1^t, ..., g_k^t, ..., g_K^t \}$ 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:

$$g_k^{t+1} = \begin{cases} g_k^t + \Delta(t, g_k^u - g_k^t) & if \quad b = 0 \\ g_k^t - \Delta(t, g_k^t - g_k^l) & if \quad b = 1 \end{cases} \tag{19}$$

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

$$\Delta(t, z) = z(1 - r^{(1-t/T)^h}) \tag{20}$$

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

2.1.4 Results

2.1.4.1 Parameter settings

The parameters of GAs are set as follows. Population size=100, crossover rate=0.9, a=0.3, h=0.5, T=200. Redundancy index threshold (L) and conflicting index threshold (T) are set as 0.75 and 0.5, respectively. The number of generations is set as 200.

2.1.4.2 Comparisons

A k-fold (k=3 in this paper) cross-validation method is adopted for algorithm comparisons. A total of 537 accident cases (1,074 parties) are randomly and equally divided into three subsets, each of which contains 179 cases (358 parties). Each model is trained and validated three times separately. The average training and validation results of various algorithms are reported in Table 3. As noted from Table 3, GRM1 performs the best, both in training and validation, with correctness rates of 78.85% and 70.21%, respectively, followed by GRM2 with training and validation correctness rates of 74.34% and 69.33%, respectively. GRM3 performs the worst with training and validation correctness rates ranging from 69.20%~71.21% and 65.84%~67.55% under various preset numbers of rules, respectively. It is worthy of noting that the correctness rate of GRM3 does not increase monotonically as anticipated as the preset number of rules gets larger. It might be partly due to the searching space being exponentially increased as the preset number of rules gets larger, making the optimal combination of rules rather difficult to be mined. In terms of the number of rules mined, the GRM1 model selects the largest number of rules (34 rules), followed by 30 rules mined by GRM3 with K=35. Even under the limitation of at most 10 rules mined (GRM3 with K=10), the correctness rates can still reach 69.20% in training and 66.46% in validation.

Table 3
The training and validation results of various GRM models

Models	Maximum number of rules allowed	Number of rules mined	Correctness rate (%)	
			Training	Validation
GRM1	100	34	78.85	70.21
GRM2	100	28	74.34	69.33
GRM3	10	10	69.20	66.46
	15	15	69.60	66.15
	20	20	70.00	67.08
	25	24	69.47	65.84
	30	28	69.23	66.87
	35	30	71.21	67.55

To get an in-depth investigation to the rules mined by the best performing model, GRM1, the fitness value, coverage ratio, predicted accuracy, redundancy index and conflicting index of a total of 34 mined rules are shown in Table 4. These rules are ranked according to fitness value f_i in Eq.(11). Note that in terms of coverage ratio, R17 can cover the largest number of parties (46% of 1,074 parties, *i.e.* 494 parties), followed by R1 (43% of 1,074 parties, *i.e.* 462 parties). In contrast, R33 and R34 are customized to rather few parties with a low coverage ratio of 1%. In general, the more explanatory variables introduced into the antecedent of a rule, the lower the coverage ratio of the rule will be.

In terms of predicted accuracy, R19 and R21 perform the best with predicted accuracy of 91%, followed by R4 and R20 with predicted accuracy of 85%. R33 and R34 have the lowest predicted accuracy of 50%. In terms of redundancy index, R7 has the highest value of 72%, which is found to be highly overlapping with R2, followed by R4 of 65% overlapping with R2. Actually, both R7 and R4 are belonging to the family rules of R2. There are still many rules having no redundancy to their previous rules. In terms of conflicting, R23 performs the worst with conflicting index as high as 42%, which highly conflicts with R3, followed by R12 with conflicting value of 21% which conflicts with R8.

Also note that a total of 17 rules (the largest number) are mined of the predicted class of “no responsibility,” followed by six rules with the predicted class of “full responsibility.” Only two rules with the predicted class of “equal responsibility” are mined.

Table 4
Combination of rules mined by the GRM1 model

Rules	Antecedent	Consequent	Fitness value	Coverage ratio	Predicted accuracy	Redundancy index	Conflicting index
Rule 1	If location="straight road" and right-of-way="yes"	Then liability="no"	0.28	0.43	0.83	0.00	0.00
Rule 2	If alcoholic use="no" and relative direction="opposing direction" and right-of-way="yes"	Then liability="no"	0.22	0.33	0.83	0.21	0.00
Rule 3	If relative direction="same direction" and right-of-way="no"	Then liability="full"	0.20	0.34	0.79	0.00	0.00
Rule 4	If type of vehicle="passenger car" and alcoholic use="no" and relative direction="opposing direction" and right-of-way ="yes"	Then liability="no"	0.16	0.22	0.85	0.65	0.00
Rule 5	If location="straight road" and type of vehicle="passenger car" and speeding="no" and right-of-way="yes"	Then liability="no"	0.15	0.23	0.81	0.55	0.00
Rule 6	If location="straight road" and foresight of the accident="no" and driver death ="no" and right-of-way ="yes"	Then liability="no"	0.14	0.22	0.82	0.48	0.00
Rule 7	If speeding="no" and relative direction="opposing direction" and right-of-way="yes"	Then liability="no"	0.11	0.37	0.65	0.72	0.00
Rule 8	If relative direction ="right adjacent direction" and right-of-way="yes"	Then liability="minor"	0.11	0.27	0.71	0.17	0.00
Rule 9	If relative direction="left adjacent direction" and right-of-way="no"	Then liability="major"	0.10	0.28	0.69	0.00	0.00
Rule 10	If relative direction="opposing direction" and right-of-way="no"	Then liability="full"	0.09	0.35	0.63	0.00	0.00
Rule 11	If location="unsignalized intersection" and right-of-way="yes"	Then liability="minor"	0.09	0.34	0.63	0.25	0.00
Rule 12	If location="signalized intersection" and speeding="no" and driver death ="no" and right-of-way ="yes"	Then liability="no"	0.09	0.19	0.73	0.32	0.21
Rule 13	If type of vehicle="passenger car" and foresight of the accident="not clear" and right-of-way ="no"	Then liability="full"	0.08	0.28	0.65	0.12	0.08
Rule 14	If speeding="no" and alcoholic use="no" and relative direction="opposing direction" and driver death ="no" and right-of-way ="yes"	Then liability="no"	0.08	0.27	0.66	0.52	0.00
Rule 15	If speeding="seriously" and relative direction="opposing direction" and right-of-way ="no"	Then liability="full"	0.07	0.21	0.67	0.45	0.00

Rule 16	If location=“unsignalized intersection” and right-of-way=“no”	Then liability=“major”	0.07	0.36	0.59	0.00	0.00
Rule 17	If speeding=“no” and foresight of the accident=“yes” and right-of-way=“yes”	Then liability=“no”	0.06	0.46	0.57	0.32	0.00
Rule 18	If alcoholic use=“no” and foresight of the accident=“not clear” and right-of-way =“yes”	Then liability=“no”	0.06	0.29	0.60	0.25	0.00
Rule 19	If foresight of the accident=“no” and relative direction=“same direction” and right-of-way =“yes”	Then liability=“no”	0.06	0.07	0.91	0.00	0.00
Rule 20	If location=“straight road” and type of vehicle =“truck” and driver death=“no” and right-of-way =“yes”	Then liability=“no”	0.06	0.08	0.85	0.21	0.00
Rule 21	If location=“straight road” and type of vehicle =“light truck” and right-of-way =“no”	Then liability=“full”	0.05	0.07	0.91	0.00	0.00
Rule 22	If foresight of the accident=“not clear” and driver death=“no” and right-of-way=“yes”	Then liability=“no”	0.04	0.27	0.57	0.69	0.00
Rule 23	If speeding =“no” and foresight of the accident=“no” and driver death=“no” and right-of-way =“no”	Then liability=“major”	0.03	0.33	0.55	0.57	0.42
Rule 24	If location=“flashlight intersection” and type of vehicle =“passenger car” and relative direction=“same direction” and right-of-way =“yes”	Then liability=“minor”	0.03	0.06	0.78	0.17	0.08
Rule 25	If type of vehicle=“light truck” and relative direction=“same direction” and driver death =“no” and right-of-way =“yes”	Then liability=“no”	0.03	0.06	0.74	0.39	0.00
Rule 26	If type of vehicle=“passenger car” and foresight of the accident=“no” and right-of-way=“yes”	Then liability=“no”	0.02	0.41	0.53	0.55	0.00
Rule 27	If foresight of the accident=“no” and relative direction=“left adjacent direction” and right-of-way =“yes”	Then liability=“minor”	0.02	0.05	0.69	0.05	0.03
Rule 28	If location=“straight road” and type of vehicle=“light truck” and speeding=“no” and right-of-way =“no”	Then liability=“minor”	0.02	0.16	0.56	0.00	0.00
Rule 29	If type of vehicle=“truck” and alcoholic use=“no” and relative direction=“opposing direction” and right-of-way =“no”	Then liability=“full”	0.01	0.03	0.70	0.21	0.00
Rule 30	If speeding =“seriously” and foresight of the accident=“yes” and right-of-way =“yes”	Then liability=“equal”	0.01	0.04	0.61	0.07	0.05
Rule 31	If type of vehicle =“passenger car” and foresight of the accident=“no” and relative direction=“opposing direction” and right-of-way =“yes”	Then liability=“no”	0.01	0.05	0.56	0.11	0.00
Rule 32	If type of vehicle=“light truck” and foresight of the accident=“yes” and right-of-way =“no”	Then liability=“major”	0.00	0.06	0.53	0.05	0.00

Rule 33	If relative direction="same direction" and driver death="yes" and right-of-way="yes"	Then liability="no"	0.00	0.01	0.50	0.01	0.00
Rule 34	If location="straight road" and type of vehicle="truck" and speeding="yes" and right-of-way ="yes"	Then liability="equal"	0.00	0.01	0.50	0.00	0.00

Table 5 further compares the results from our proposed GRM1 model and from the ANN and DA models developed by Chiou (2006). In terms of correctness rate, ANN2 model (case-based) with 10 hidden neurons performs the best both in training and validation, followed by ANN1 model (party-based) with 10 hidden neurons, ANN1 (12-10-1), in training and the proposed GRM1 model in validation. The DA model still performs the worst. In consideration of model's accuracy and complexity, Chiou (2006) finally selected ANN1 (12-5-1) as the best performed model according to SBC. However, it is hard to identify the number of parameters of genetic algorithms, thus SBC of the proposed GRM1 model is not computed. Nonetheless, it is worth noting that the correctness rates of training and validation of GRM1 model are both higher than ANN1 (12-5-1).

Table 5
Comparisons with ANN and DA models developed by Chiou (2006)

Models	*H*	Training		Validation	
		Correctness rate (%)	SBC	Correctness rate (%)	SBC
GRM1	–	78.85	–	70.21	–
ANN1	5	78.17	-0.82*	66.10	0.13*
	10	82.43	-0.35	68.75	1.04
	15	80.72	0.38	55.27	2.39
ANN2	5	78.92	0.55	64.14	2.70
	10	85.72*	2.56	77.91*	5.73
	15	70.59	4.70	65.25	9.84
DA	–	59.05	-0.82*	54.09	0.13*

Note: SBC stands for Schwarz's Bayesian Criterion. SBC=ln(*MSE*)+*P*ln(*N*)/*N*, where, *MSE* is mean squared error, *P* is the number of parameters. * indicates the best performing model in terms of corresponding criterion. *H* is the number of hidden neurons.

To gain in-depth investigation on the validation results of different models, the number of parties with degrees of liabilities predicted by DA, ANN1 (12-5-1), and GRM 1 are reported in Tables 6~8, respectively. Note that the DA model has the highest correctness rate of 66.90% for the category of real y = 5 and the lowest correctness rate of 5.00% for the category of real y = 3, suggesting that DA model performs rather poorly in the cases of equal liability. In contrast, ANN1 (12-5-1) model can achieve over 60% of correctness rate almost for all categories, except for the category of real y = 4. Besides, the degrees of liabilities even incorrectly predicted by ANN do not deviate over one degree from their real ones; except for a total of 19 (1.77%) parties which are deviated two degrees. The proposed GRM1 model can accurately predict the liability for the categories of y = 1 or 5 with correctness rates of 92.25% and 91.90%. Even for the categories of y = 2 or 4, the proposed model can still reach about 60% of correctness rate. However, the proposed model is unable to predict the liability for the category of y = 3. The correctness rate of that category is only 8.33%. Nonetheless, the degrees of liabilities, even incorrectly predicted by GRM, do not deviate over one degree from their real ones; except for a total of 3 parties which are deviated two degrees. Besides, the proposed GRM1 model tends to produce much more predicted classes at both extremes: y=1 or y=5 than at the middle, *i.e.* y=3.

Table 6
The number of parties with degrees of liabilities predicted by the DA model

Real *y*	Predicted *y*					Total
	1	2	3	4	5	
1	157 (55.28)	98 (34.51)	20 (7.04)	9 (3.17)	0 (0.00)	284 (100.00)
2	86 (38.57)	125 (56.05)	0 (0.00)	11 (4.93)	1 (0.45)	223 (100.00)
3	18 (30.00)	12 (20.00)	3 (5.00)	12 (20.00)	15 (25.00)	60 (100.00)
4	2 (0.90)	36 (16.14)	0 (0.00)	106 (47.53)	79 (35.43)	223 (100.00)
5	0 (0.00)	3 (1.06)	30 (10.56)	61 (21.48)	190 (66.90)	284 (100.00)
Total	263	274	53	199	285	1074

Note: The percentages are given in the parentheses.

Table 7
The number of parties with degrees of liabilities predicted by the ANN1 model

Real *y*	Predicted *y*					Total
	1	2	3	4	5	
1	202 (71.13)	69 (24.30)	13 (4.58)	0 (0.00)	0 (0.00)	284 (100.00)
2	54 (24.22)	139 (62.33)	30 (13.45)	0 (0.00)	0 (0.00)	223 (100.00)
3	0 (0.00)	11 (18.33)	41 (68.33)	8 (13.33)	0 (0.00)	60 (100.00)
4	0 (0.00)	4 (1.79)	44 (19.73)	107 (47.98)	68 (30.49)	223 (100.00)
5	0 (0.00)	0 (0.00)	2 (0.70)	61 (21.48)	221 (77.82)	284 (100.00)
Total	256	223	130	176	289	1074

Note: The percentages are given in the parentheses.

Table 8
The number of parties with degrees of liabilities predicted by the GRM1 model

Real *y*	Predicted *y*					Total
	1	2	3	4	5	
1	262 (92.25)	20 (7.04)	2 (0.71)	0 (0.00)	0 (0.00)	284 (100.00)
2	88 (39.46)	133 (59.64)	2 (0.90)	0 (0.00)	0 (0.00)	223 (100.00)
3	0 (0.00)	24 (40.00)	5 (8.33)	31 (51.67)	0 (0.00)	60 (100.00)
4	0 (0.00)	3 (1.35)	0 (0.00)	137 (61.43)	83 (37.22)	223 (100.00)
5	0 (0.00)	0 (0.00)	0 (0.00)	23 (8.10)	261 (91.90)	284 (100.00)
Total	350	180	9	191	344	1074

Note: The percentages are given in the parentheses.

2.1.4.3 Discussions

The proposed GRM1 models may not exhibit remarkably higher correctness rate than all of ANN models developed by Chiou (2006); however, the proposed models can generate meaningful rules for further examination and demonstration, making the GRM-based expert system much more understandable than the black-box ANN approach. Consequently, decision makers might feel more confident in using this GRM-based expert system. Besides, with rules mined, the accident appraisal knowledge can be clearly displayed and used to train junior members. And, it also offers a post-optimization mechanism which can further fine-tune the mining rules by interviewing the accident appraisal experts.

To further investigate the rules mined in Table 4, only eight out of twelve explanatory variables appeared in at least one rule, including location, type of vehicle, speeding, alcoholic use, foresight of the accident, relative direction, driver death, and right-of-way. Particularly, all of the rules introduce the variable of right-of-way, explaining its importance in accident appraisal. Furthermore, most of the rules with right-of-way="yes" will also lead to the consequents of liability="no responsibility" or "minor responsibility." In contrast, the rules with right-of-way="no", then the liabilities are most likely to be "major responsibility" or "full responsibility." As such, one may

argue that the accident appraisal can be conducted solely depending upon right-of-way. However, in doing so, the correctness rate is only 48.65%, which is even lower than that of the DA model.

Unlike the right-of-way being a decisive factor, some variables appearing in the rules are more likely to act like an environmental factor, such as location and relative direction, to determine whether the accident situation is clear or ambiguous. In a relatively clearer situation, such as location="straight road," or relative direction="same direction" or "opposing direction", the liability degree tends to be overwhelmingly assessed as either "no responsibility" or "full responsibility" solely depending on right-of-way, such as the rules of R1, R2, R3, R5, and R6 etc. In contrast, in a relatively more ambiguous situation, such as location="unsignalized intersection" or "flashlight intersection" or relative direction="left adjacent direction" or "right adjacent direction," then the liability degree are more conservatively assessed as "minor responsibility" and "major responsibility," such as the rules of R8, R9, R11, and R16 etc.

On the other hand, some variables operate like an incremental factor to further raise or alleviate the liability degree accompanying with the ownership of right-of-way. These variables are violation variables, including speeding and alcoholic use. Taking R30 for instance, the liability degree is suggested as "equal responsibility," when the party was seriously speeding even with the ownership of right-of-way.

High redundancy relationship can still be found among rules mined. Taking R4 for instance, it belongs to the family rules of R2. It is said that R4 is more specific than R2 by further specifying the type of vehicle. Due to the high redundancy between these two rules, deleting any one of them may not seriously deteriorate the correctness rate, if the number of rules is strictly limited.

In comparing the degree of liability predicted with different models, the proposed GRM model performs worse in predicting the category of liability="equal responsibility," because very few related rules (with consequent part of liability="equal") have been mined. Two reasons could be identified. First, the number of parties with liability="equal responsibility" in the dataset is only approximately one fourth of other degrees, thus it is difficult to learn representative rules from limited parties. Second, since the right-of-way is very decisive to liability assessment, all predicted results are clearly divided into two distinct categories: no (minor) and full (major) degrees, leaving no much room for middle degree.

Four explanatory variables: type of road, direction, crash spot, and LAC, are not included in any rule of a total of 34 rules. It can be explained that they are not key factors to the accident appraisal. However, it might also be possible that these variables are categorized into too many classes, *e.g.* crash spot (8 classes) and LAC (12 classes). In general, GRM tends not to choose a variable with too many classes, because the number of cases or parties belonging to each class will be small and lower the coverage ratio. Thus, it would be very helpful to collect more cases with equal responsibility for training or to approximately merge some classes into fewer categories.

2.1.5 Concluding remarks

This paper employs genetic rule mining (RGM) to develop accident appraisal expert systems by discovering knowledge from historical appraisal cases, which can not only accurately predict the liability degrees of involved parties but also demonstrate comprehensive appraisal rules and further provide the flexibility of post-adjustment of mined rules. Three GRM models based on the Michigan approach (GRM1 and GRM2) and the Pittsburgh approach (GRM3) have been developed in this study. To effectively mine rules based on the Michigan approach, a novel two-stage rule selection procedure is proposed. The first stage is to rank the chromosomes survived from selection, crossover, and mutation operations according to a fitness function composed by coverage ratio, predicted accuracy and predicted error. The second stage is then to filter the ranked chromosomes depending on whether their redundancy and conflicting indices are larger than the preset thresholds.

For training and validating the proposed three models, a total of 537 two-car crash accident cases (1074 parties) are randomly and equally divided into three subsets, each of which contains

179 cases (358 parties). Each model is trained and validated three times separately. The results show that the GRM1 model, party-based Michigan approach, performs the best with training and validation correctness rates of 78.85% and 70.21%, respectively. Comparisons with the ANN1, ANN2, and DA models proposed by Chiou (2006) also show that the proposed GRM1 model can achieve the accuracy level of ANN models, but more importantly, the GRM1 model can generate a comprehensive combination of rules. Through an in-depth investigation on a total of 34 rules mined by GRM1 model, some underlying accident appraisal knowledge can be extracted. First, right-of-way is found to be a decisive factor which not only shows in every rule mined but also determines most of the liability degree as two distinct categories: no (or minor) responsibility vs. full (or major) responsibility depending upon which party owns the right-of-way. Second, some environmental variables, such as location and relative direction, appear in rules to determine how clear the accident situation was, which then lead the liability degree to be either an overwhelming result: no responsibility vs. full responsibility or to be a conservative result: minor responsibility vs. major responsibility. Third, some violation variables, such as speeding or alcohol use, are then used to add or relieve liability degree which has been assessed based on other facts.

Several directions can be identified for future studies. First, the encoding method for the Pittsburgh approach proposed in this paper is rather lengthy, as a result that the case-based Pittsburgh model can not be considered. A more compact and efficient encoding method for Pittsburgh approach deserves further studies. Besides, the proposed two-stage chromosome ranking and filtering procedure of the Michigan approach is somewhat inefficient because of ranking process involved. A more efficient procedure deserves further explorations. Although the proposed GRM1 model can accurately predict the liability degrees of $y = 1$ or 5 with correctness rate over 90%, and can still predict at a satisfactory correctness rate for the liability degrees of $y = 2$ or 4 (about 60%), it performs rather poorly in predicting the liability degree of $y = 3$. Since the number of training samples with such a liability degree is only one-fourth of samples with other degrees, we believe that it can further enhance the predictive capability of proposed models by collecting more such accident appraisal cases. It is also essential to fine-tune the rules mined by interviewing senior accident appraisal experts through an interactive manner. Last but not least, the fact that some variables do not appear in any rule might partly contribute to their values being classified into too many categories, leading to a rather small share of cases or parties covered in each category. To properly merge their categories without losing too much meaningful knowledge is also worthy of exploration.

2.2 事故分析

2.2.1 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, *et al.*, 1996; Shanker and Mannering, 1996; Dissanayake

et al., 2002, Al-Ghamdi, 2002; Delen, *et al*., 2002; Tay and Rifaat, 2007; Sze and Wong, 2007).

Although statistic models are the most commonly used methods in the context of crash data analysis either collectively or individually, most of them have their own assumptions and complexity in model estimation and interpretation. Once the assumptions were violated, the model could lead to erroneous estimation results. Especially for the individual approach, most of the variables explaining the individual crashes are categorical, such as driver gender, road type, lighting condition, violation, weather condition, and severity degree, etc. It is difficult to develop parametric statistical models based upon such categorical data. Therefore, a number of distribution-free methods, particularly for dealing with classification and prediction problems such as decision tree (Chang and Chen, 2005; Chang and Wang, 2006) and artificial neural network (Chiou, 2006; Delen *et al.*, 2006), were adopted. 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 a black-box characterized method 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 problems of conflicts and redundancy among the rules mined, our proposed GMR model will account for the conflicts and redundancy, in addition to conventional objectives: coverage ratio and predictive accuracy.

2.2.2 Data

The crash data was 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. Each 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 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 numbers of 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 of 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.

2.2.3 Genetic rule mining

Genetic rule mining (GMR), which can automatically learn of comprehensive rules from available dataset and toward decision support, has been shown as a useful tool 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.

2.2.3.1 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 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.

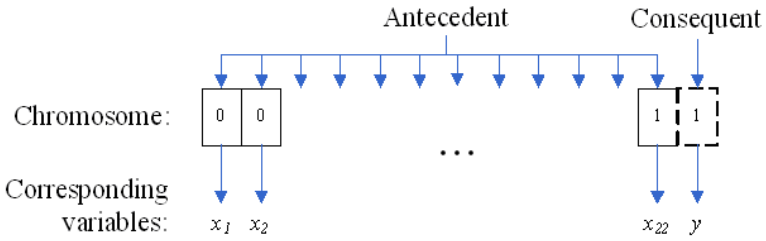


Figure 1 Encoding method of the proposed GRM model

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 equals to 23. Each gene will then take one of the categorical values of the corresponding variable. Since 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 a value of 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} x_{22}={0, 1,…, 6} and Then y=2." That is, any case satisfies any one of the offspring rules will certainly also satisfy their parent rule. Generally, the more variable present in the antecedent part (taking none 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 most accurately predict the liability degree based upon these twelve explanatory variables. The total number of potential rules equals to $3\times3\times4\times4\times4\times5\times5\times5\times5\times5\times5\times5\times5\times6\times6\times6\times6\times6\times6\times6\times8\times7\times4 = 1.058\times10^{16}$. Obviously, it is barely possible to compare all rule combinations through a total enumeration approach.

2.2.3.2 Fitness Function

An individual chromosome, a rule, with a higher fitness function value has a higher probability to be selected to reproduce offspring. Obviously, the role of the fitness function is to evaluate the quality of the rule numerically. In doing so, three common factors to be taken into account are 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:

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

2.2.3.3 Genetic Operators

Because the genes in our GRM 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:

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

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 a 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.

2.2.3.4 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 cannot be avoided that two or more rules with different predicted classes might still 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 numbers of fired rules. As such, the correctness rate of the model can be computed as the number of correctly predicted cases divided by the total number of cases.

2.2.4 Results and discussions

2.2.4.1 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=1000.

Table 2 shows the final selected rules along with its corresponding performance indices. Note that a total of 19 rules are selected with a descending order according to f_i. In terms of fitness value (f_i), the top seven rules have remarkably higher values than the rest of twelve rules, suggesting that it is promising to use only the top seven rules to conduct the prediction. In terms of coverage ratio (CR_i), R7 can explain 828 cases, followed by R2 (633 cases) and R6 (597 cases). In contrast, some rules cover only very few cases, such as R14 (3 cases) and R18 (4 cases). In terms of predictive accuracy (PA_i), R1 has the highest predictive accuracy of 0.930, followed by R14 (0.667), and R19 the least (0.094).

The importance of variable can be identified by the number of variables presenting in rules. In this regard, x_5 (obstacle) is the most important variable which appears at 13 rules (appearance rate 68.42%), followed by x_1 (surface condition) with appearance rate 47.37%. Four variables are not shown in any rule, which are x_6 (lighting condition), x_9 (marking), x_{16} (location), and x_{17} (vehicle type), indicating their insignificance to crash severity. There are four rules associated with A1 crash, six rules with A2 crash, and nine rules with A3 crash.

Table 2 Combination of rules mined by GRM 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}	x_{22}	y	f_i	CR_i	PA_i
R1		1			1						3		1										1	54.997	71	0.930
R2	1				1						1	1											3	37.604	633	0.395
R3	1	1			1														1				3	31.436	533	0.394
R4					1		2				1	1						1					3	28.689	238	0.500
R5							2																3	23.711	447	0.380
R6	1											1							1				2	23.652	597	0.345
R7			1		1							1											1	21.196	828	0.298
R8				3																			3	9.077	195	0.364
R9	1			1						1				1									3	8.952	180	0.372
R10			2		1		2					1											3	4.134	40	0.475
R11	1	1	1		1									4									3	3.901	90	0.356
R12												1		3								4	2	3.720	49	0.429
R13		1			1										1								2	3.675	80	0.363
R14	1										3		1										2	0.857	3	0.667
R15	1			2	1							1						1					1	0.439	40	0.225
R16	1		1		1		4						1							3			3	0.393	29	0.241
R17	1				2																		2	0.125	31	0.161
R18					1		4	4				1									5		1	0.060	4	0.250
R19			1		1						3								1				2	0.025	32	0.094
n	9	4	5	3	13	0	5	1	0	1	5	8	2	3	1	0	0	2	3	1	1	1	-	-	-	-

Note: *n* is the number of appearance of the variable in the selected 19 rules.

Table 3 gives the distribution of cases with degree of severity predicted by GRM model and with real degree of severity. As shown in Table 3, in the training dataset, the proposed GRM model can actually predict A3 crash with a correct rate of 96.82%, followed by A1 crash (correct rate 73.99%) and A2 (correct rate 64.64%). The overall correct rate of the proposed GRM model achieves 78.50%. In the validation dataset, the overall correct rate slightly has decreased to 74.16%.

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

Datasets	Real severity	Predicted severity			Total
		A1	A2	A3	
Training	A1	256 (73.99%)	84 (24.28%)	6 (1.73%)	346 (100.00)
	A2	51 (14.78%)	223 (64.64%)	71 (20.58%)	345 (100.00)
	A3	4 (1.16%)	7 (2.02%)	335 (96.82%)	346 (100.00)
	Total	311	314	412	1037
Validation	A1	105 (70.95%)	37 (25.00%)	6 (4.05%)	148 (100.00)
	A2	15 (10.07%)	96 (64.43%)	38 (25.50%)	149 (100.00)
	A3	5 (3.38%)	14 (9.46%)	129 (87.16%)	148 (100.00)
	Total	125	147	173	445

Note: The percentages are given in the parentheses.

2.2.5 Comparisons

To prove the performance of our proposed GRM model, a decision tree (DT) model is also used to mine the rules explaining the same crash dataset. 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 for predicting A2 crash (correct rates in training and validation are 78.84% and 77.18%, respectively) than the proposed GRM model. However, the DT model performs much worse than the proposed GRM 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 61.25%, respectively, which are much lower than those of the proposed GRM model. Thus, the performance of the proposed GRM model has been proven.

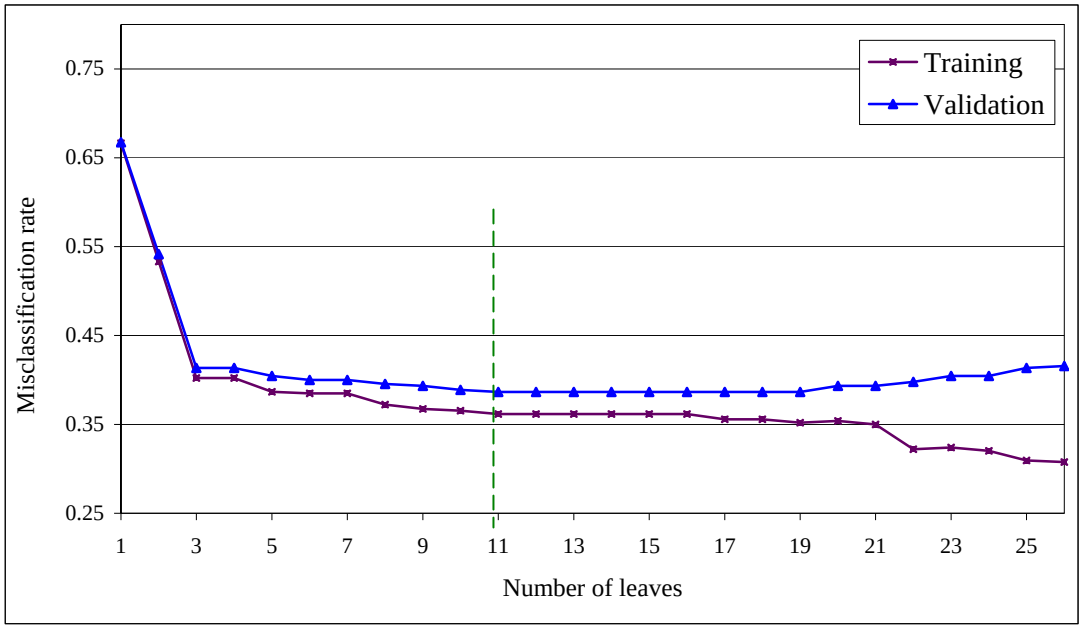


Figure 2 Learning process of the DT model

Table 4 Number of cases with degree of severity predicted by DT

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.

A total of 11 rules are generated by the DT model as follows. There are a total of 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

2.2.6 Conclusion

This paper employs individual approach to identify contributing factors to crash severity by developing a novel genetic rule mining (GRM) model. To avoid over-mining problem caused by unevenly distributed cases across degrees of severity, identical numbers of A1-type, A2-type and

A3-type of crash cases are selected from 2003-2007 Taiwan freeway accidents dataset. A total of 19 rules have been mined which can achieve an overall correct rate of 78.50% in training and 74.16% in validation, respectively, which have demonstrated much higher correctness than the conventional decision tree model. The performance of the proposed GRM model has been proven. According to the mined rules, x_5 (obstacle) and x_1 (surface condition) are two key factors contributing to crash severity.

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 have 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 simplify the model complexity, various performance indices may be integrated into an overall fitness function; as such, a multi-objective GRM model deserves further elaboration. Last but not least, more comparisons can be made to other commonly used methods (e.g., logistic regression model, artificial neural network) to demonstrate the superiority of our proposed GRM model.

三、計畫成果自評

本計畫為三年期計畫。其中，本期中報告已完成第一年期之研究內容，另外也完成部分第二研究年期之內容。此兩年期所列之主要工作項目，預期之研究成果如下：

（一）第一個研究年期：

1.彙整事故鑑定案例及挑選候選變數

本研究將蒐集整理臺灣省覆議鑑定委員會鑑定案件，並從中挑選地區鑑定委員會與臺灣省覆議鑑定委員會鑑定結果一致（各地區鑑定會判定肇事當事人有責任者，臺灣省覆議鑑定會亦判定其相同肇事責任）之案件資料，以作為模式學習與應用之案例。為進一步了解影響鑑定之變數與肇事責任判定的關係，本研究將以交叉分析進行篩選影響變數，以挑選對肇事責任判定較具顯著影響的關鍵變數作為模式之輸入變數。

2.建立基因規則探勘模式

本研究將分別以密西根方式或匹茲堡方式進行染色體之編碼設計，再比較此兩種方式之優劣差異。以密西根方式而言，可將一染色體編碼如下：

3.建立事故責任鑑定專家系統

分別利用所建構之基因規則探勘模式應用於事故責任鑑定案例。在規則學習與驗證上，本研究採用三次交叉驗證方式（threefold cross-validation），將所有案例分為三部份。分別以第一及第二部份案例作為訓練資料，以第三部份案例作為驗證資料。再改以第一及第三部份作為訓練資料，以第二部份作為驗證資料。最後，改以第二及第三部份作為訓練資料，以第一部份作為驗證資料，三次平均可得訓練及驗證績效平均值。

4.模式績效之比較分析

本研究將除進行密西根及匹茲堡方式之比較外，並將同時與邱裕鈞、方守潔（民 93）所建立之判別分析及類神經網路模式之訓練與驗證結果，進行比較。

5.推理規則之產生與詮釋

經由比較分析後，可依據表現較佳之基因規則探勘模式之挑選規則，進行分析及詮釋，並加以整理列表。透過與資深覆議會鑑定委員之訪談，以了解模式所挑選之規則是否符合鑑

定委員進行鑑定之推理邏輯，以作為本模式尋優方向與績效評估之調校或可應用性之驗證。

6.參數敏感度分析

由於遺傳演算法之尋優績效可能會受其參數設定值之影響。因此，為了解本模式是否具有穩定性（robustness），將針對不同參數設定組合（包括交配率、突變率、族群數等）之尋優結果進行比較分析，以驗證本模式之穩定性，並提出參數設定之建議數值。

（二）第二個研究年期

1.蒐集高速公路事故資料並篩選重要解釋變數

高速公路事故資料分為 A1（死亡事故）、A2（受傷事故）、A3（財損事故）三大類，以民國 94 年為例，全年共計 13,661 件。其中，A1 事故 118 件，A2 事故 965 件，A3 事故 12,570 件。相關變數包括：事故發生時間地點、當時天候狀況資料、當地道路幾何條件資料、事故類型、主要肇事原因、傷亡狀況、交通管制狀況、駕駛人行為與違規狀況等。將利用交叉分析表方式，先作顯著變數之初步篩選。

2.建立高速公路事故分析與預測模式

分別利用所建構之基因規則探勘模式（第一個研究年期）及螞蟻規則探勘模式應用於高速公路事故分析與預測案例。在規則學習與驗證上，本研究採用三次交叉驗證方式（threefold cross-validation），將所有案例分為三部份。分別以第一及第二部份案例作為訓練資料，以第三部份案例作為驗證資料。再改以第一及第三部份作為訓練資料，以第二部份作為驗證資料。最後，改以第二及第三部份作為訓練資料，以第一部份作為驗證資料，三次平均可得訓練及驗證績效平均值。

3.模式績效之比較分析

本研究將除與基因規則探勘模式進行比較外，並將同時與羅吉斯迴歸、判別分析及類神經網路模式之訓練與驗證結果，進行比較。

4.推理規則之產生與詮釋

經由比較分析後，可依據表現較佳之規則探勘模式所挑選規則，進行分析及詮釋，並加以整理列表。以深入了解各環境變數群、交通管制變數群、駕駛人行為變數群間對事故嚴重性之聯合效果關係，並據以研提改善策略。

上述第一年期之預期研究成果已順利達成，而本年期也進一步提出第二年期之部分研究成果。為下一年度之研究奠定良好基礎。此外，本計畫之主要成果已分別發表國際研討會 1 篇文章[19]，並已改寫投稿學術期刊中[18, 19]。此外，本計畫亦用以指導一名博士生進行論文寫作[20]。

四、參考文獻

1. Booker, L.B., Goldberg, D.E., & Holland, J.H. (1989). Classifier systems and genetic algorithms. *Artificial Intelligence, 40*, 235-282.
2. Chiou, Y.C. (2006). An artificial neural network-based expert system for the accident appraisal of two-car crash accidents. *Accidents Analysis & Prevention, 38(4),* 777-785.
3. Chiou, Y.C., & Lan, L.W. (2005). Genetic fuzzy logic controller: An iterative evolution algorithm with new encoding method. *Fuzzy Sets and Systems, 152(3),* 617-635.
4. Dehuri, S., & Mall, R. (2006). Predictive and comprehensible rule discovery using a

multi-objective genetic algorithm. *Knowledge-Based Systems*, *19,* 413-421.
5. Dehuri, S., Patnaik, S., Ghosh, A., & Mall, R. (2008). Application of elitist multi-objective genetic algorithm for classification rule generation. *Applied Soft Computing, 8,* 477-487.
6. DeJong, K. (1988). Learning with genetic algorithms: An overview. *Machine Learning*, 3(3), 121-138.
7. Fedelis, M.V. (2000). Discovering comprehensive classification rules with a genetic algorithm. *Proceeding of Congress on Evolutionary Computation (CEC'2000)*. La Jolla Marriott, San Diego, CA, USA, July 16-19.
8. Freitas, A.A. (1999). On rule interestingness measures. *Knowledge-Based Systems, 12,* 309-315.
9. Ghosh, A., Nath, B. (2004). Multi-objective rule mining using genetic algorithms, *Information Sciences, 163,* 123-133.
10. Goldberg, D.E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley.
11. Herrera, F., Lozano, M., & Verdegay, J.L. (1995). Tuning fuzzy logic controllers by genetic algorithms. *International Journal of Approximate Reasoning, 12,* 299-315.
12. Herrera, F., Lozano, M., & Verdegay, J.L. (1998). A learning process for fuzzy control rules using genetic algorithms. *Fuzzy Sets and Systems, 100,* 143-158.
13. Holland, J.H., & Reitman, J.S. (1978). Cognitive systems based on adaptive algorithms. In Waterman, D.A., & Hayes-Roth, F. (Eds.), *Pattern-Directed Inference Systems*. Academic Press.
14. Lekova, A., Mikhailov, L., Boyadjiev, D., & Nabout, A. (1998). Redundant fuzzy rules exclusion by genetic algorithms. *Fuzzy Sets and Systems, 100,* 235-243.
15. Michalewicz, Z. (1992). *Genetic algorithms + data structures = evolution programs*, Springer, Berlin.
16. Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. in: G. Piatetsky-Shapiro, W.J. Frawley (Eds.), *Knowledge Discovery in Databases,* AAAI, 229.
17. Smith, S.F. (1983). Flexible learning of problem solving heuristics through adaptive search. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 422-425. Morgan Kaufmann.
18. Chiou, Y.C. and Chen, C.Y., (2009) “Development of two-car crash accident appraisal expert system with genetic rule mining,” submitted to Journal of Safety Research (Under review)
19. Chiou, Y.C., Lan, W.L. and Chen, W.B. (2009) “Contributory factors to crash severity in Taiwan freeways: Genetic mining rule approach,” submitted to Journal of Eastern Asia Society for Transportation Studies. (Under review)
20. 陳文斌，Identifying contributory factors to crash severity in Taiwan freeways by genetic rough set rule mining，交通大學交通運輸研究所，博士論文（進行中），民國98年。