ELSEVIER

Accident Analysis and Prevention 39 (2007) 629–637

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Rough set approach for accident chains exploration

Jinn-Tsai Wong [a,*], Yi-Shih Chung [b,1]

[a] Institute of Traffic and Transportation, National Chiao Tung University, 4F, 114 Chung Hsiao W. Road, Sector 1, Taipei 100, Taiwan
[b] Institute of Traffic and Transportation, National Chiao Tung University, 3F, 114 Chung Hsiao W. Road, Sector 1, Taipei 100, Taiwan

Received 8 August 2006; received in revised form 20 September 2006; accepted 14 October 2006

Abstract

This paper presents a novel non-parametric methodology – rough set theory – for accident occurrence exploration. The rough set theory allows researchers to analyze accidents in multiple dimensions and to model accident occurrence as factor chains. Factor chains are composed of driver characteristics, trip characteristics, driver behavior and environment factors that imply typical accident occurrence. A real-world database (2003 Taiwan single auto-vehicle accidents) is used as an example to demonstrate the proposed approach. The results show that although most accident patterns are unique, some accident patterns are significant and worth noting. Student drivers who are young and less experienced exhibit a relatively high possibility of being involved in off-road accidents on roads with a speed limit between 51 and 79 km/h under normal driving circumstances. Notably, for bump-into-facility accidents, wet surface is a distinctive environmental factor.

Keywords: Rough set theory; Accident characteristics; Accident chain

1. Introduction

Statistical models, such as linear regression, logistic regression, Poisson, negative binomial or zero-inflated count models (Al-Ghamdi, 2002; Chandraratna et al., 2006; Kim and Kim, 2003; Lee et al., 2002; Lord et al., 2005; Maher and Summersgill, 1996; Miaou and Lum, 1993; Oh et al., 2004; Ozkan and Lajunen, 2006), have been widely applied in analyzing the relationships between factors and accidents. Contributing factors were successfully identified. However, due to the extremely complicated relationships among factors and the restrictions of data and methods (Elvik, 2003, 2006; Hauer, 1997), it has been difficult for researchers to properly build links among affecting factors and accident outcomes. The causes of an accident have usually been described with the closest-to-accident factors. Researchers, however, have recently tended to analyze an accident more thoroughly – not only the accident itself but also the activities and factors prior to and subsequent to the accident. Some accident patterns were found to be preventable not by correcting driving behaviors but by adjusting behaviors prior to driving (Eby et al., 2000; Simoes, 2003). An accident, therefore, may not occur if one or more undesirable activities in this process were removed (Baker and Ross, 1961; Fleury and Brenac, 2001; Reason, 1997).

Modeling accidents as chains or combinations of factors and consequences has become an alternative way to understand the process of accidents. The derived chains imply causality between factors and consequences, and are usually called causal chains (Elvik, 2003) or scenarios (Fleury and Brenac, 2001). The sequence does not necessarily represent the actual accident process but rather the logical and temporal links between factors and consequences (Elvik, 2003). The generating process of accidents can be either analyzed by extracting the circumstances of certain accident consequences (types or severity) or by comparing the differences of circumstances between consequences. The causality between factors and consequences is interpreted by the derived outcomes, which are usually represented as combinations of factors, trees or rules. Several techniques have been applied in this context including hierarchical ascendant clustering (Berg et al., 2004; Laflamme and Eilert-Petersson, 1997), entropy classification methods (Strnad et al., 1997; Vorko and Jović, 2000), classification trees and neural networks (Chang and Wang, 2006; Delen et al., 2006; Karlaftis and Golias, 2002; Sohn and Shin, 2001). In this paper, however, the aim is to introduce a relatively new method called rough set theory as a complement to the complicated accident chain analysis.

* Corresponding author. Tel.: +886 2 2349 4959; fax: +886 2 2349 4953.
E-mail addresses: jtwong@mail.nctu.edu.tw (J.-T. Wong), yschung.tt93g@nctu.edu.tw (Y.-S. Chung).
[1] Tel.: +886 2 2349 4995; fax: +886 2 2349 4953.

0001-4575/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2006.10.009

Although it is impossible to model complete causal chains for each accident because of the data limitation, the results derived from rough set can be treated as pseudo-causal chains which explain the indispensable factors involved in the occurrence of a certain type of accident. The word *pseudo* reflects the unique essence of each accident where current databases cannot contain all the possible factors. Fortunately, the connections between accidents and affecting factors can be tested by alternating the set of factors being considered. Meanwhile, the characteristics of critical accident chains, i.e. the derived rules from rough set, can be explored by analyzing various linkages between factors being considered and the resulted factor combinations.

Rough set theory was proposed by Pawlak (1982). It is a relatively new classification method in data-mining field and has been shown to be a useful mathematical tool for exploring data patterns (Greco et al., 2001). Rough set theory is an extension of set theory; it can effectively handle discrete variables with multilevel categories. It is non-parametric and avoids issues such as membership functions in fuzzy theory. Instead of null hypothesis of significance testing, rough set theory provides an alternative way to evaluate the importance of factors. This may partially relieve Hauer's concern (2004) about confusion of non-significant factors in statistical sense with unimportant factors in common sense. Furthermore, the introduced rough set theory can classify accidents into groups with similar properties by considering multiple dimensions that help reduce the unobserved heterogeneity (Karlaftis and Tarko, 1998). Thus, we believe that rough set theory has the potential to be a complementary method for analyzing relationships among factors and accidents taking into considerations the process of accident occurrence.

The rest of this paper is organized as follows. Rough set theory is introduced and the research framework is proposed in Sections 2 and 3, respectively. Thereafter, a real-data set is adopted to demonstrate rough set and the proposed framework in Section 4. Discussion follows in Section 5 and conclusions are drawn in Section 6.

2. Rough set theory for exploring accident occurrence

Let U represent the universe, a finite set of objects, and P denote a set of condition attributes, i.e. affecting factors for the occurrence of accidents. For x, $y \in U$, we say that x and y are indiscernible by the set of condition attributes P if $\rho(x, q) = \rho(y, q)$ for every $q \in P$ where $\rho(x, q)$ denotes the information function. A set that has objects within it that are indiscernible by the set of condition attributes P is called a P-elementary set. The family of all elementary sets is denoted by P^*. 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 P-lower approximation of a set of objects Y ($Y \subseteq U$), denoted by $\underline{P}Y$, and the P-upper approximation of Y, denoted by $\bar{P}Y$, are defined as

$$\underline{P}Y = \bigcup X \quad \{X \in P^* \text{ and } X \subseteq Y\}$$

$$\bar{P}Y = \bigcup X \quad \{X \in P^* \text{ and } X \cap Y \neq \emptyset\}$$

The objects belonging to the set of lower approximation are those definitely definable by the elementary sets, since objects in $\underline{P}Y$ can be fully identified by the elementary sets in P^*. 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 P^*.

As illustrated in Table 1, five cases are characterized with three condition attributes: *driver's age*, *vehicle type* and *climate*, and one decision attribute: *accident type*. 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 off-road accident type is described with the lower approximation set, {2}, and the upper approximation set, {1,2,3}. Similarly, the concept of the rollover 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 percentage of the associated objects definable with the specified condition attributes. This can be defined as follows:

$$\pi_P(Y) = \frac{\text{card}(\underline{P}Y)}{\text{card}(\bar{P}Y)}$$

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; i.e. more accident cases of this accident type are discernible by the elementary sets generated by the specified condition attributes. This implies that the associated accident patterns do exist unambiguously. Following the Table 1 example, the accuracy of approximation for the off-road type is 0.33(=1/3) while that for the rollover type is 0.50(=2/4). This implies the rollover type can be defined more unambiguously than the off-road type 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. $Y_i \cap Y_j = \emptyset$, $\forall i, j \leq n$ $i \neq j$ and $\bigcup_{i=1}^{n} Y_i = U$. Y_i are called classes of X. The P-lower approximation and P-upper approximation of X are represented by sets $\underline{P}X = \{\underline{P}Y_1, \underline{P}Y_2, \ldots, \underline{P}Y_n\}$ and $\bar{P}X = \{\bar{P}Y_1, \bar{P}Y_2, \ldots, \bar{P}Y_n\}$, respectively. Quality of approximation of classification X by a set of attributes can be defined as follows:

$$\gamma_P(X) = \frac{\sum_{i=1}^{n} \text{card}(\underline{P}Y_i)}{\text{card}(U)}$$

Table 1
Example of accident cases with describing features

Case	Driver's age	Vehicle type	Climate	Accident type
1	Young	Motorcycle	Sunny	Off-road
2	Old	Automobile	Sunny	Off-road
3	Young	Motorcycle	Sunny	Rollover
4	Middle-aged	Motorcycle	Sunny	Rollover
5	Middle-aged	Automobile	Rainy	Rollover

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. This implies that the accident chains for all accident types 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). This implies that with the provided three condition attributes 60% of the cases can be unambiguously defined.

To recognize further the details of accident chains, *rules* need to be extracted. A rule, representing the critical chain 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 show up since such accidents have never happened before. A rule exists if and only if at least one such accident exists. Many rule generation algorithms have been proposed in recent years (Greco et al., 2001), but it is beyond this paper's scope to discuss those algorithms. This paper simply applies the most frequently used algorithm – minimum covering – to generate rules. Its aim is to generate the minimum number as well as the shortest length of rules to cover all accidents.

In the next section, a research framework will be proposed to show how to utilize the information provided by an accident database to investigate the links among affecting factors and accident outcomes.

3. Research framework

An accident database usually consists of three types of information: person, vehicle and accident characteristics. These data were the observable and recordable information that described the process of accident occurrence. While the sequence of some factors in the process of accident occurrence cannot be determined, other factors can be. For example, mode choice and trip time must have been decided before driving. Consequently, we can distinguish recorded information into four sets: driver's characteristics, trip characteristics, behavior and environment factors, and accident information. The first three are the condition attributes and the last refers to the decision attribute. Driver's characteristics usually include age, education, race and gender. Trip characteristics describe the feature of a trip such as trip time, mode choice and trip purpose. Behavior and environment (B&E) factors consist of a driver's on-road behavior, road environment, and natural environment such as cell phone use and light condition. Accident is the output of interest such as different accident types or severity. Thereafter, a typical accident process can be described sequentially – deciding trip-making characteristics, driving on roads, and finally, being involved in an accident – while driver characteristics influence all those activities in different ways. This process is shown in Fig. 1.

In order to examine the contribution of each set of variables in distinguishing accident types, the factor sets are analyzed with rough set by different combinations rather than all at once. Seven approaches linking these factors sequentially to accidents can be demonstrated:

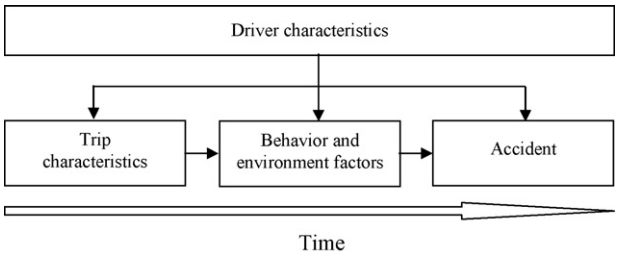

Fig. 1. Factor structure.

- *Approach 1*. Driver characteristics → accident;
- *Approach 2*. Trip characteristics → accident;
- *Approach 3*. B&E factors → accident;
- *Approach 4*. Driver characteristics → trip characteristics → accident;
- *Approach 5*. Driver characteristics → B&E factors → accident;
- *Approach 6*. Trip characteristics → B&E factors → accident;
- *Approach 7*. Driver characteristics → trip characteristics → B&E factors → accident.

All the approaches, except Approach 7, have been well examined and some interesting results have been found (Sümer, 2003 for Approaches 1, 3 and 5; Eby et al., 2000; Simoes, 2003 for Approach 4; Lenne et al., 1997; Maycock, 1995; Vollrath et al., 2002 for Approach 6). However, no research demonstrated differences among those approaches and the associated implications. Different approaches imply that different sets of condition attributes are loaded to describe the occurrence of accidents and then different families of elementary sets are generated. Some accident types may be well defined by a certain collection of condition attributes while others may not be. By comparing the loaded attributes and classification results, we can distinguish the essential differences between accident types. Characteristics of factor chains can be obtained which help understand the nature of occurrence of accidents.

4. Empirical study for accident chains exploration

In order to demonstrate the effectiveness of rough set theory and the proposed approach, an empirical study is presented in the following.

4.1. Data

Taiwan 2003 single auto-vehicle (SAV) accident data is chosen to demonstrate the feasibility and usefulness of rough set theory and the proposed framework in accident chain analyses. Single auto-vehicle accidents are those in which only a single vehicle is involved. The total number of SAV accidents, excluding invalid cases, was 2316. The collected attributes and their corresponding categories are summarized in Table 2. Accident type is chosen as the decision attribute while the other attributes are considered as condition attributes. A popular rough set software, ROSE2 (rough set data explorer), is used in this

Table 2
Attribute and category

Dimension	Attribute	Category
Driver characteristics (condition attribute)	Age	Under (<18), young (18–35), middle-aged (36–55), old (>55)
	Gender	Male, female
	License type	Regular, occupational, military, other
	License condition	Valid, invalid, unknown
	Occupation	Student, working people, no job, other, unknown
Trip characteristics (condition attribute)	Trip purpose	Work, school, social, shop, sightseeing, business, other, unknown
	Trip time	Morning peak (07:00–09:00 h), day offpeak (09:00–16:00 h), afternoon peak (16:00–19:00 h), night offpeak (19:00–23:00 h), midnight to daybreak (23:00–07:00 h)
Behavior and environment factors (condition attribute)	Protect equipment use	Use, no use, unknown
	Cell phone use	Use, no use, unknown
	Drinking condition	Drinking, not drinking, other
	Road type	Highway, other
	Speed limit	50−, 51–79, 80+
	Road shape	Intersection, segment, ramp, other
	Pavement material	Asphalt, other, no pavement
	Surface deficiency	Normal, other (e.g. holes, soft, and so on)
	Surface status	Dry, wet, other
	Obstruction[a]	Yes, no (within 15 m)
	Sight distance	Good, poor (based on road design speed)
	Signal type	Regular, flash, no signal
	Signal condition	Normal, abnormal, no signal
	Direction divided facility	Island, marking, none
	Roadside marking	Yes, no
	Climate	Sunny or cloudy, rainy, other
	Light condition	With light, no light
Accident (decision attribute)	Accident type	Bump into bridge or architecture (198)[b]
		Bump into road facility (1564)
		Bump into non-fixed object (17)
		Bump into work zone (21)
		Off-road (297)
		Rollover (93)
		Other (126)

[a] Regular road facility excluded.
[b] Sample size of the accident type.

paper where LEM2 (Grzymala-Busse, 1992; Grzymala-Busse and Werbrouck, 1998) is embedded to generate a minimum rule set covering all objects.

4.2. Results of rough set analysis

The results of rough set analysis consist of five parts: rule generation, quality of approximation, rule validation, rule description and significance of condition attributes. The results of the first three parts are summarized in Table 3.

4.2.1. Rule generation

As shown in Table 3, the number of rules generated increases with the completeness of the specified condition attributes. Since all the condition attributes are categorical variables, the incorporation of any additional condition attribute with n categories would expand the possible classifications n times. However, while the quality of approximation is much enhanced, the number of rules does not increase proportionally but only with limited growth. This implies that the condition attributes included are valid enough to classify the accident types and that some patterns do exist for the SAV accidents in Taiwan rather than all SAV accidents being regarded as unique.

4.2.2. Quality of approximation

The accuracy of approximation for rollover and bump-into-non-fixed object accidents is extremely low, except when all condition attributes are included. However, the accuracy of approximation for the bump-into-bridge accidents, off-road accidents, and other accident types can be increased to 30–40% if B&E factors are combined with either driver characteristics or trip characteristics. This can be raised to 70% or even 80% if all condition attributes are included. Roughly speaking, bump-into-facility and work zone are the most definable accident types, while bump-into-bridge, off-road, and other accident types are moderately definable accident types, and rollover and bump-into-non-fixed object are the least definable accident types.

The quality of classification is proportional to the completeness of selected attributes. Approach 7 shows the highest quality, while Approach 2 shows the lowest. B&E factors show the most important attributes for the quality of classification partly due

Table 3
Rough set results

Approach		Accident type	Generated rules	Accuracy (%)	Quality of classification (%)	Hit rate (%)	Overall hit rate (%)
		Bridge		0.19		4.55	
		Facility		2.26		5.05	
	D[a]	Non-fixed obj.		0.00		11.76	
1	↓	Work	104	0.00	3.02	23.81	6.30
	A	Off-road		0.51		4.04	
		Rollover		0.11		12.90	
		Other		0.50		21.43	
		Bridge		0.00		0.00	
		Facility		0.26		1.73	
	T	Non-fixed obj.		0.00		23.53	
2	↓	Work	38	0.00	0.26	23.81	4.62
	A	Off-road		0.00		11.11	
		Rollover		0.00		27.96	
		Other		0.00		9.52	
		Bridge		7.52		21.21	
		Facility		31.59		27.88	
	B	Non-fixed obj.		1.59		29.41	
3	↓	Work	508	9.66	38.69	23.81	25.60
	A	Off-road		7.88		21.21	
		Rollover		2.34		24.73	
		Other		4.96		15.08	
		Bridge		1.68		20.20	
	D	Facility		16.47		21.93	
	↓	Non-fixed obj.		0.97		0.00	
4	T	Work	474	0.76	20.16	23.81	20.60
	↓	Off-road		3.38		19.53	
	A	Rollover		1.77		18.28	
		Other		1.31		11.11	
		Bridge		39.89		16.67	
	D	Facility		67.78		53.45	
	↓	Non-fixed obj.		19.64		0.00	
5	B	Work	766	79.17	74.65	9.52	42.01
	↓	Off-road		41.39		22.90	
	A	Rollover		17.39		21.51	
		Other		30.22		11.11	
		Bridge		31.29		19.19	
	T	Facility		64.05		49.36	
	↓	Non-fixed obj.		8.43		0.00	
6	B	Work	787	45.95	70.68	14.29	39.21
	↓	Off-road		33.96		21.55	
	A	Rollover		18.38		18.28	
		Other		21.11		11.11	
	D	Bridge		74.68		12.63	
	↓	Facility		90.57		69.69	
	T	Non-fixed obj.		41.94		5.88	
7	↓	Work	808	100.00	92.88	23.81	51.38
	B	Off-road		80.65		17.51	
	↓	Rollover		66.39		9.68	
	A	Other		69.81		6.35	

[a] D: driver characteristics; T: trip characteristics; B: behavior and environment factors; A: accidents.

to their wide coverage of affecting factors, which are also proximal factors. Each dimension alone (Approaches 1–3) does not yield a good quality of classification. If at least two dimensions are combined, the quality of classification is much enhanced. For example, the quality of classification for B&E alone is 38.69%. However, it is raised to 70.68% by merely combining it with trip characteristics in which only two more attributes are included.

These results suggest that accidents should not be resolved by single factor, but by a chain of factors. Previous countermeasures focused mostly on B&E proximal factors. It is effective; however, to further improve road safety, all factors associated in the factor chain may need to be taken into serious consideration. Furthermore, neglecting factors in a chain may result in rather different stories and blur the interactions among accident features.

4.2.3. Rule validation

The 10-fold cross-validation technique is used to conduct validation test of classification results. The hit rate, i.e. the percentage of correct prediction, for the bump-into-facility accidents can be improved by up to 70% when all condition attributes are considered. On the other hand, the hit rates for the remaining accident types all range from 0 to 20 or 30%. This suggests that the occurrence of a bump-into-facility accident may follow similar paths and is more predictable. But for other accident types, the rules generated from their training cases may not be representative since their occurrences are mostly random.

The higher the quality of approximation, the higher the overall hit rate and the hit rate for the bump-into-facility accidents. Yet, the bump-into-bridge and bump-into-non-fixed object accidents show the highest hit rate in Approach 3, which consists of B&E proximal factors only and reveals the unexpected and random characteristics of these kinds of accidents. Its hit rate becomes lower if other condition attributes are included. These results suggest that except for the bump-into-facility accidents where more information is useful, different accident types have their corresponding useful condition attributes. For example, the condition attributes of driver characteristics are useful for the bump-into-work zone and the other accident types, and those of trip characteristics are useful for rollover accidents. All these results are helpful for devising adequate countermeasures.

The classification results show that most of the bump-into-bridge, bump-into-facility, off-road and rollover accidents are assigned to the bump-into-facility accident type and least into the bump-into-non-fixed and bump-into-work zone accident types. This suggests that, while most accidents are associated with some critical condition attributes which lead to the similar classification pattern, bump-into-non-fixed and bump-into-work zone accidents are related to very distinctive characteristics. This also implies that some similarities may exist in the occurrence of the bump-into-bridge, bump-into-facility, off-road and rollover types since they are all related to road geometry and driving environments. These similarities are the reasons for the low hit rates for the bump-into-bridge and off-road accident types, since they can be easily assigned to the bump-into-facility accidents due to the fact that the sample size for the bump-into-facility accident type outweighs theirs. As a consequence, more rules associated with the occurrence of the bump-into-facility accident type are generated and dominate the classification pattern. On the other hand, the remaining accident types, such as the bump-into-non-fixed object, are more closely related to driver characteristics and are relatively unique.

4.2.4. Description of significant rules

Rules are generated from the accident database by rough set theory, and the significant rules for each accident type are shown in Table 4. The rule strength – the number of accident cases matching the rule – for most accident types is small except for the bump-into-facility type. The highest strength for most types is about 3 or 4. This shows the uniqueness of those accident types, especially, the infrequent and stochastic occurrences of the bump-into-non-fixed objects. Interestingly, the derived factor chain shows that a drinking driver without regular license exhibits a relatively high possibility of being involved in bump-into-non-fixed object accidents on a secondary road without roadside marking and light.

The most significant rule for the bump-into-work zone suggests that there is a relatively high risk when a driver approaches work zone on a road with speed limit less than 50 km/h around midnight. This information suggests that more effective and sufficient work zone traffic controls should be installed, particularly in the dark work zone on those secondary roads. The rule reflects the fact that, to save cost, it is often the case that safety measures are not properly implemented, especially on rural secondary roads.

For rollover accidents, two significant rules describe young working people who are driving during off-peak period as being more likely involved in the rollover accidents, probably due to the low traffic and high speed.

Four significant rules for the bump-into-bridge accidents describe two conditions: drinking driving under normal road environment and sober drivers under abnormal road environment. Specific deficiencies exist on both conditions for this accident type. This shows the necessity for the government to prevent this type of accident by improving the road environment or raising the penalties for drinking driving.

The derived factor chain for off-road accidents shows that student drivers who are young and less experienced exhibit a relatively high possibility of being involved in off-road accidents. This result echoes the graduated licensing scheme currently existing in many countries (Simpson, 2003). Moreover, the factor chain shows that the corresponding driving environment is normal, i.e. no particularly unfavorable factors such as drinking driving or poor sight distance appear on the chain. Since other driving groups such as working people do not show similar accident patterns as off-road accident type, the government should seriously consider educating student drivers to enhance their situational awareness of driving environment and reduce their risk-driving behavior on roads.

The rule with the highest strength goes to bump-into-facility accidents. It describes 35 employed sober drivers rather than students driving on an island-divided road segment where the surface was wet and there were no obstructions within 15 m. The wet surface denotes lower friction on road surfaces that increase the difficulty of handling vehicles. Meanwhile, drivers generally might slow down their driving speed to maintain vehicles at an "acceptable" speed. Therefore, the extremely high supporting evidence may imply that those drivers overestimated their driving skills and underestimated the risk of the decrease in surface friction.

4.2.5. Significance of condition attributes

The significance of condition attributes is measured by their presence on the derived rules. When a condition attribute shows up more frequently in the rules, it is more likely being used to describe the occurrence of accidents and hence is more significant in distinguishing accident types. The presence of a condition attribute is represented with presence percentage which is calculated by summing up its presence in each rule weighted with cases of the associated rule divided by total cases. Here, only

Table 4
Description of significant rules

Accident type	Rule description[a]
Bump into facility (35)[b]	Driver: working people Behavior: not drinking Environment: road segment; median island; wet surface; no obstruction within 15 m
Off-road (7)	Driver: regular license; student Environment: speed limit 51–79; median marking; with roadside marking; with light
	Driver: middle-aged; working people Behavior: drinking Environment: speed limit less than 50; collision position rather than intersection, segment and ramp; with roadside marking
Bump into bridge or architecture (4)	Behavior: drinking; cell phone use unknown Environment: flash signal; no roadside marking; dry surface
	Driver: young; working people Trip: other trip purpose; between midnight and daybreak Behavior: not drinking Environment: no signal; median marking; with roadside marking; with light; poor sight distance
	Behavior: not drinking Environment: collision position rather than intersection, segment and ramp; pavement rather than asphalt; no directional-divided facility; no roadside marking; no obstruction within 15 m
Bump into work zone (4)	Driver: male; regular license type; unknown occupation Trip: during midnight to daybreak Environment: speed limit less than 50; asphalt pavement; no signal; obstruction within 15 m
Rollover (3)	Driver: young; working people Trip: social trip; night offpeak Behavior: not drinking Environment: median marking
	Driver: young; male; regular license type; working people Trip: day offpeak Environment: speed limit less than 50; regular signal
Bump into non-fixed object (2)	Driver: other license type Behavior: drinking; cell phone use unknown Environment: speed limit less than 50; no roadside marking; no light

[a] Please refer to Table 2 for the details of condition attributes.
[b] The value represents the rule strength.

the rules derived from Approach 7 are adopted in the calculation since Approach 7 shows the most satisfactory performance. Moreover, since condition attributes with more categories tend to distinguish accident types more effectively, comparisons are made on those with same number of categories. As shown in Fig. 2, gender, roadside marking and light condition; speed limit, road shape and directional divided facility; age, occupation, trip time and drinking condition are those attributes with a relatively higher presence percentage among all condition attributes with two, three and four or more categories, respectively.

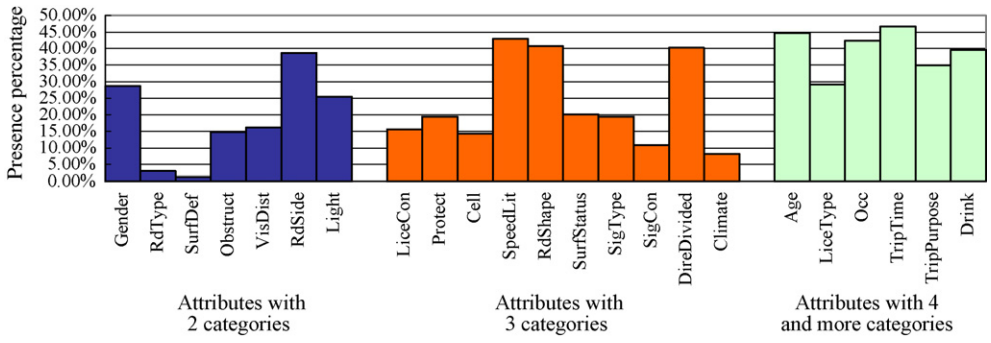

Fig. 2. Presence percentage of condition attributes.

5. Discussion

Taking advantages of rough set, this research implemented the idea that the occurrence of an accident is a series of errors or mishandling. The illustrated case shows that it is feasible to apply rough set theory to analyze the links among affecting factors and accident types. The proposed factor structure can be easily transformed and extended based on an analyst's knowledge and his/her on-hand accident databases. Any factor structures can be tested by similar steps proposed in this research. In addition, a large number of condition attributes were included without any prior judgments except when being grouped with respect to the temporal and logical sequence of the occurrence of an accident. A condition attribute was dropped only when the removal did not have any impact on defining accident types. In our empirical study, only one redundant condition attribute (pavement material) was found when all the attributes were included. This procedure differs from conventional statistical approaches where non-significant attributes are usually immediately dropped and are sometimes claimed to have no impact on the occurrence of an accident.

Rules generated from rough sets provide fruitful information describing conditions under which certain type of accidents may occur. For example, as mentioned in the previous section, the most significant rule for the bump-into-work zone suggests that there is a relatively high risk when a driver approaches work zone on a road with speed limit less than 50 km/h around midnight. When it comes to employment of the modern ITS technologies (FHWA, 2006), specific warning messages could be devised and sent to the drivers conforming to this particular scenario; consequently, the potential accidents could be prevented. In short, the derived rules have the potential to distribute the right information to the right drivers at the right time for them to be able to act properly.

On the other hand, hundreds of rules were generated in the end, which makes it difficult for analysts to conclude which rules or accident patterns are the most significant. This result may partly come from the fact that some accident types, such as the bump-into-non-fixed object accidents or rollover accidents, are so stochastic and unique, and partly from the lack of detailed information about drivers' characteristics in the database that hinder the possibility of more effectively recognizing accident characteristics. Despite the fact that these accident types are the least definable and the least classifiable, some protective measures still can be implemented to reduce the accident possibility and severity such as preventing animals crossing roads or increasing the strength of the vehicle roof. On the other hand, the most definable and recognizable accident type – the bump-into-facility accidents – is regarded as being preventable. In addition, the bump-into-bridge and off-road accidents showing similar classification patterns as the bump-into-facility accidents, are also expected to be preventable.

In order to find representative rules for occurrence of those avoidable accident types, more advanced rough set models, such as the hybrid approach combining rough set with genetic programming (McKee and Lensberg, 2002), can be adopted in future research. However, for the low-performing (unpredictable) accident types which are highly related to driver characteristics and unpredictable environment conditions (i.e. non-fixed objects), more related data need to be collected for further study. Meanwhile, instead of preventing accidents, measures for reducing the negative effects of those unpredictable accidents may be more effective and are worth investigating.

The estimation results showed that the accuracy of approximation, the quality of approximation and the hit rates could be dramatically enhanced by considering at least two sets of condition attributes while the inclusion of overall condition attributes generally gave the most satisfactory quality of classification. This suggests that collecting more detailed data on some specialties rather than aimlessly increasing survey items is more effective. Nonetheless, additional attributes are welcomed and could be collected and examined by testing their redundancy and their effect on the accuracy of approximation, quality of approximation as well as hit rates to determine whether they are worthwhile.

6. Conclusions

In this paper, we have proposed a research framework integrated with rough set to analyze the information provided by accident databases. The existence of accident patterns for Taiwan 2003 single auto-vehicle accidents and their major differences and similarities are demonstrated. Overall, we have succeeded in applying rough set in mining some information from an accident database that provides a lot of messages for effective accident-prevention decision-making.

This research is a new attempt to apply rough set as a complementary tool for accident analyses. A lot of information is still embedded in the derived rules that might provide useful knowledge for researchers and analysts and requires further exploration. Meanwhile, different accident types such as multi-vehicle accidents can be examined with similar approaches. Advanced models, however, should be considered in the future to improve and to address the issues related to performance of rule extraction and case validation.

Acknowledgments

The authors would like to thank the anonymous referees for their helpful comments and suggestions, and the National Science Council of Taiwan for providing the research grant (NSC 94-2211-E-009-030).

References

Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accid. Anal. Prev. 34 (6), 729–741.

Baker, J.S., Ross, H.L., 1961. Concepts and classification of traffic accident causes (part 1). Int. Road Safety Traffic Rev. 9 (31), 11–18.

Berg, H.Y., Gregersen, N.P., Laflamme, L., 2004. Typical patterns in road-traffic accidents during training—an explorative Swedish national study. Accid. Anal. Prev. 36 (4), 603–608.

Chandraratna, S., Stamatiadis, N., Stromberg, A., 2006. Crash involvement of drivers with multiple crashes. Accid. Anal. Prev. 38 (3), 532–541.

Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid. Anal. Prev. 38 (5), 1019–1027.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38 (3), 434–444.

Eby, D., Shope, J.T., Molnar, L.J., Vivoda, J.M., Fordyce, T.A., 2000. 2000 Improvement of Older Driver Safety through Self Evaluation: The Development of a Self-Evaluation Instrument. UMTRI Report-2000-04. Transportation Research Institute, University of Michigan.

Elvik, R., 2003. Assessing the validity of road safety evaluation studies by analyzing causal chains. Accid. Anal. Prev. 35 (5), 741–748.

Elvik, R., 2006. Laws of accident causation. Accid. Anal. Prev. 38 (4), 742–747.

Fleury, D., Brenac, T., 2001. Accident prototypical scenarios, a tool for road safety research and diagnostic studies. Accid. Anal. Prev. 33 (2), 267–276.

FHWA, 2006. Safety applications of intelligent transportation systems in Europe and Japan. FHWA-PL-06-001. Federal Highway Administration, Department of Transportation, Washington, DC.

Greco, S., Matarazzo, B., Slowinski, R., 2001. Rough sets theory for multicriteria decision analysis. Eur. J. Operat. Res. 129 (1), 1–47.

Grzymala-Busse, J.W., 1992. LERS—A System for Learning from Examples Based on Rough Sets. Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publisher, Dordrecht.

Grzymala-Busse, J.W., Werbrouck, P., 1998. On the best search method in the LEM1 and LEM2 algorithms. In: Orlowska, E. (Ed.), Incomplete Information: Rough Set Analysis. Physica-Verlag, Heidelberg, New York, pp. 75–91.

Hauer, E., 1997. Observational before-after studies in road safety. In: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Pergamon Press, Oxford.

Hauer, E., 2004. The harm done by tests of significance. Accid. Anal. Prev. 36 (3), 357–365.

Karlaftis, M.G., Golias, I., 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. Accid. Anal. Prev. 34 (3), 357–365.

Karlaftis, M.G., Tarko, A.P., 1998. Heterogeneity considerations in accident modeling. Accid. Anal. Prev. 30 (4), 425–433.

Kim, S., Kim, K., 2003. Personal, temporal and spatial characteristics of seriously injured crash-involved seat belt non-users in Hawaii. Accid. Anal. Prev. 35 (1), 121–130.

Laflamme, L., Eilert-Petersson, E., 1997. School-injury patterns: a tool for safety planning at the school and community levels. Accid. Anal. Prev. 30 (2), 277–283.

Lee, A.H., Stevenson, M.R., Wang, K., Yau, K.W., 2002. Modeling young driver motor vehicle crashes: data with extra zeros. Accid. Anal. Prev. 34 (4), 515–521.

Lenne, M.G., Triggs, T.J., Redman, J.R., 1997. Time of day variations in driving performance. Accid. Anal. Prev. 29 (4), 431–437.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. 37 (1), 35–46.

Maher, M.J., Summersgill, I., 1996. A comprehensive methodology for the fitting of predictive accident models. Accid. Anal. Prev. 28 (3), 281–296.

Maycock, G., 1995. Driver Sleepiness as a Factor in Car and HGV Accidents, TRL Report. Transport Research Library, Crowthorne.

McKee, T.E., Lensberg, T., 2002. Genetic programming and rough sets: a hybrid approach to bankruptcy classification. Eur. J. Operat. Res. 138 (2), 436–451.

Miaou, S.P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design. Accid. Anal. Prev. 25 (6), 689–709.

Oh, J., Washington, S., Choi, K., 2004. Development of accident prediction models for rural highway intersections. Transp. Res. Rec. 1897, 18–27.

Ozkan, T., Lajunen, T., 2006. What causes the differences in driving between young men and women? The effects of gender roles and sex on young drivers' driving behaviour and self-assessment of skills. Transp. Res. Part F 9 (4), 269–277.

Pawlak, Z., 1982. Rough sets. Int. J. Comput. Inform. Sci. 11 (5), 341–356.

Reason, J., 1997. Managing the Risks of Organizational Accidents. Ashgate, Aldershot.

Simoes, A., 2003. The cognitive training needs of older drivers. Recherche Transp. Securite 79, 145–155.

Simpson, H.M., 2003. The evolution and effectiveness of graduated licensing. J. Safety Res. 34 (1), 25–34.

Sohn, S.Y., Shin, H., 2001. Pattern recognition for road traffic accident severity in Korea. Ergonomics 44 (1), 107–117.

Strnad, M., Jović, F., Vorko, A., Kovacic, L., Toth, D., 1997. Young child injury analysis by the classification entropy method. Accid. Anal. Prev. 30 (5), 689–695.

Sümer, N., 2003. Personality and behavioral predictors of traffic accidents: testing a contextual mediated model. Accid. Anal. Prev. 35 (6), 949–964.

Vorko, A., Jović, F., 2000. Multiple attribute entropy classification of school-aged injuries. Accid. Anal. Prev. 32 (3), 445–454.

Vollrath, M., Meilinger, T., Kruger, H., 2002. How the presence of passengers influences the risk of a collision with another vehicle. Accid. Anal. Prev. 34 (5), 649–654.