Applying Data Mining to Assess Crash Risk on Curves

Samantha Chen,1, Andry Rakotonirainy,1 Seng Wai Loke2
1 Queensland University of Technology, CARRS-Q, Queensland
2 La Trobe University, Department of Computer Science and Computer Engineering Melbourne
email: s2.chen@qut.edu.au, r.andry@qut.edu.au, S.Loke@latrobe.edu.au

Abstract

The wide range of contributing factors and circumstances surrounding crashes on road curves suggest that no single intervention can prevent these crashes. This paper presents a novel methodology, based on data mining techniques, to identify contributing factors and the relationship between them. It identifies contributing factors that influence the risk of a crash. Incident records, described using free text, from a large insurance company were analysed with rough set theory. Rough set theory was used to discover dependencies among data, and reasons using the vague, uncertain and imprecise information that characterised the insurance dataset.

The results show that male drivers, who are between 50 and 59 years old, driving during evening peak hours are involved with a collision, had a lowest crash risk. Drivers between 25 and 29 years old, driving from around midnight to 6 am and in a new car has the highest risk. The analysis of the most significant contributing factors on curves suggests that drivers with driving experience of 25 to 42 years, who are driving a new vehicle have the highest crash cost risk, characterised by the vehicle running off the road and hitting a tree.

This research complements existing statistically based tools approach to analyse road crashes. Our data mining approach is supported with proven theory and will allow road safety practitioners to effectively understand the dependencies between contributing factors and the crash type with the view to designing tailored countermeasures.

Keywords

Data mining, text mining, rough set analysis, crash risk, relationships, road curve.

Introduction

Road curves are a common feature of road infrastructure; however, the consequences of crashes that occurred at road curves could be more serious than on straight roads. In Queensland, approximately 30% of the crashes occurred on road curves [1] and 34% of the crashes are fatal or require hospitalisation [2]. Reports from Queensland Transport state that the fatalities rate on road curves is 2.5 times higher than crashes on straight roads [2].

Several studies have examined this issue, and a range of interventions were proposed. Most interventions are designed after analysing the contributing factors with the view to reduce the driver’s exposure to such factors. Crash risk is derived from the likelihood of being involved in a crash, primarily focussing on the drivers’ exposure to contributing factors. The contributing factors relate to the environment (e.g road), driver or vehicle. Examples of these interventions include road signs, warnings, and media campaigns to educate drivers.

Crash risk can be defined modelled in number of ways [3][4]. It can be the number of crash per km travelled (e.g number of crashes per 100,000 km), population, (e.g. number of crash per 500,000 habitant), location (e.g. black spot), speed or road user (e.g. cyclists), and driver impairments (e.g. fatigue or BAC). However, our notion of crash risk is calculated using the vehicle’s crash cost. The repair cost was obtained from insurance claims. We hypothesize that the insurance claim cost is correlated to crash severity. The more severe the property damage is, the higher the repair cost is.
Most existing interventions determine the crash risk as the statistical probability of a crash. The most significant factors contributing to a crash are identified using statistical methods. The contributing factors are mainly identified post-crash by police attending the crash scene. In Queensland, the Police are required to attend the crash scene only when the cost of the damage exceeds $2,000 or human injury occurred. The police officers complete a form according to their subjective understanding of the crash scene. The report is then turned into a crash report for Queensland Transport where the most significant contributing factors are identified.

Unfortunately, existing crash reports such as the one used in Queensland, do not provide detailed descriptions of crashes such as correlation between crash cost and contributing factors and the relationships between contributing factors. This prevents us from a thorough analysis of crash risk and factors.

The aim of this paper is to use data mining techniques to identify the contributing factors to curve related crashes, and then investigate the relationship between factors. It identifies contributing factors that influence the risk of a crash.

Data mining is also known as data or knowledge discovery. It is a process which analyses large volumes of data from different points of view to discover hidden correlations, patterns, trends and dependencies in a dataset. Rough set theory [5] is the formal foundation of the data mining techniques used in this paper. Data mining is not a new technique; however, to our knowledge applying this technique in crash risk assessment is novel. Identifying contributing factors with text mining technique is innovative as most, if not all existing crash analysis identify contributing factors with statistical methods.

Data mining is a process that extracts knowledge by analysing data to discover hidden patterns and dependencies in the database [6][7]. Road safety can be improved with the application of data mining techniques. Data mining techniques can be used to predict a driver’s behaviour so that unsafe actions can be rectified [8]. Researchers have performed traffic studies and investigated a method to predict the occurrence of a crash. Pande and Abdel-Aty [9] applied data mining techniques to predict rear-end crashes on highways and warn drivers about potential crashes 5-10 minutes prior to the crash. The techniques are used to identify and classify different categories of crashes and conditions that make rear-end crashes more likely. The classification results are then used to define a prediction model. The model is used in real-time with the use of loop detectors to assess the crash risk. This study shows that data mining techniques can be used to identify the causes of the crash and at the same time to predict and warn drivers of unsafe actions in specific situations.

The followings are several systems currently employed. VEDAS is data mining system that uses an on-board data stream mining and management systems [10]. VEDAS is a mobile and distributed data stream mining system for real-time vehicle monitoring. VEDAS reports emerging patterns to the fleet managers at a central control station over low-bandwidth wireless network connection. The drawback of VEDAS is that it does not have the situation awareness feature to capture contextual information of on-road conditions to improve its’ response accuracy. SAWUR is an Advanced Driving Assistance System (ADAS) that incorporates and uses ubiquitous data mining (UDM) to analyse contextual information related to driver behaviour, environment, driver profile and condition of the car in real-time.

To our knowledge, the use of data mining techniques to assess crash risk from natural text description has never been attempted.

Existing studies [11][12] have studied the relationships between contributing factors; however, the findings only relate one factor to another. This can only provide limited understanding of the interactions between multiple factors. Thus, there is a need to identify the relationships between more factors and especially those for crashes on road curves.
Methods

In this paper, we analysed the Queensland crash reports from an insurance company to identify contributing factors according to crash risk or cost. The insurance reports feature claim reports, claim cost, day of crash, age of the driver, as well as free text description of the incident completed by the policy holder. A claim record could be for less than AUD$2,500, as opposed to the Queensland crash report. The use of such database provides more insight into the characteristics of the crash which could not be otherwise obtained from the Queensland Transport database.

The insurance data set consisted of crash claim records that occurred from the year 2003 to 2006. There are 65,536 records. The data features information about the driver, vehicle, along with a free form text description of the crash. It also contains attributes such as gender, driver age, alcohol consumption, manufactured year of vehicle, time or date of crash, crash description (unstructured data), type of crash (rear or front), as well as the number of parties involved and crash costs.

Our method identifies contributing factors from the incident descriptions using text mining technique. The incident descriptions are composed of blocks of free form text written by the policy holder. The free text details the crash in natural language (English) from the policy holders’ perspective. Traditional data mining technique have only been used to analyse numerical data. Thus, this project has used a different method and used text mining to analyse the text description.

The incident descriptions are used as input for the text mining process. This is achieved with a text mining module in SAS called the Text miner. The mining module clustered the data based on the WARD algorithm. Text miner identifies the keywords along with the frequency count. The frequency count is used to identify the most frequently used keywords among the free text descriptions. The keywords with the highest count are identified as the contributing factors. The contributing factors are compared to a set of keywords from a set of non-curve related crash records in order to validate that they are factors related to road curves only.

Once the contributing factors are identified, the relationship between these factors was established. This is achieved with rough set analysis which is able to find the relationship and a minimum set of contributing factors to represent the data. The minimum contributing factor can be used to indicate the most significant contributing factor, while the relationship is indicated with the possible combinations among the contributing factors. The relationships are described as rules. The rules are classified into one of the risk levels defined. The risk levels are made up of the following five levels: (1) lowest, (2) low, (3) medium, (4) high, (5) highest.

In addition, the rules are examined thoroughly to locate any redundant or useless rules. For this research, the rules are filtered and selected based on the confidence level. Confidence can be also known as the strength, which is used to measure the quality of the rules obtained [13]. This can be done using filters available in Rosetta. Rosetta [14] is the software which implements rough set analysis.

Results

The text mining generates a list of words that appeared in the data and the associated frequencies used. The most frequent contributing factors are: rock, wet, lost control, kangaroo, dog, gravel, dirt, left turn, right turn, oncoming traffic, narrow, tree, oil, wrong direction, sharp, truck, overtake, and pole. The types of crashes associated with the keywords are: collide or collision, spin, veer, swerve and slide. Note that an initial analysis comparing contribution factors related to crash on curves and crashes outside road curves were conducted. The results showed that the keywords were significantly different.

Furthermore, the newly identified contributing factors complement those previously identified in the Queensland Transport Report [2]. The following factors, right-hand turn, narrow road, truck and avoiding kangaroo or dog are not present as contributing factors in the Queensland Transport database.

The rough set analysis process generated 2577 rules. As it is a large number, we reduced the number based on the quality. Quality is assessed with the strength of the rule. In Rosetta, the support count is the
measure of the strength of the rule [14][15]. The relative strength is computed by dividing the support count over by the total attributes and multiplying by 100. The higher the support count, the higher the strength. Thus the rules with higher strength are selected. Rules with low or small quality of strength are not considered as such rules cannot provide accurate prediction of crash circumstances.

From the rule selection process, the rules with the stronger strength or highest relative support for each risk level are shown in Table 1. The rules represent a possible combination or relationship between the contributing factors. The risk levels are represented in each column and the contributing factors for the rules are represented in rows. The contributing factors are classified into driver-related factors, road and environment-related factors, driver error, vehicle-related factors, and the crash type. The last row represents the relative support of the rule in each risk level. The rules for each risk level are read with an invisible AND between each row. For example, the rule for lowest risk consists of the following related contributing factors: male AND age between 50 to 59 AND with driving experience from 25 to 42 years AND driving at 1600 to 1900 hour AND normal road surface AND no object AND no kerb AND no lost control AND no alcohol AND vehicle is manufactured between 1991 to 2000 year AND no run-off-road crash AND collision AND no hit object.

Table 1: The strongest rules for each risk level.

<table>
<thead>
<tr>
<th>Risk level</th>
<th>Lowest</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Highest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norm cost* (AU$)</td>
<td>0.00-2.49</td>
<td>2.50-33.00</td>
<td>33.10-71.00</td>
<td>71.10-113.41</td>
<td>113.42-215.22</td>
</tr>
<tr>
<td>Risk description</td>
<td>Driver related</td>
<td>Road &amp; Environment</td>
<td>Driver Error</td>
<td>Vehicle Year</td>
<td>Crash type</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>Time</td>
<td>Road Surface</td>
<td>Object</td>
<td>Kerb</td>
</tr>
<tr>
<td></td>
<td>50-59</td>
<td>16-19</td>
<td>Norm</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>25-29</td>
<td>12-16</td>
<td>Norm</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>30-39</td>
<td>19-24</td>
<td>Norm</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>50-59</td>
<td>6-9</td>
<td>Wet</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>25-29</td>
<td>12-6</td>
<td>Norm</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative support (%)</td>
<td>13.79</td>
<td>27.58</td>
<td>3.45</td>
<td>3.45</td>
<td>3.45</td>
</tr>
</tbody>
</table>

*Norm cost = normalised cost

The rules are generated using rough set theory and Table 1 presents the strong rules with highest support count for each risk level. Strong rules are rules that are evaluated from an appropriate combination of support and accuracy characteristics [16]. The highest support count is 8 and the relative strength is 27.58%. This means that this rule supports 27.58% of the data.

Besides knowing the relationship, rough set is able to identify the minimum or significant contributing factors. Figure 1 presents the distribution of the significant contributing factors for each risk level. The significant contributing factors contains factors that are not listed in QT reports such as tree, lost control, the age of the vehicle and the driving experience. This contributes new knowledge to the existing list of contributing factors.
Figure 1: The distribution of significant contributing factors among each risk level.

Legend: vehage(2) is the year the vehicle is manufactured from year 1991 to 2000 inclusive. vehage(1) is new vehicle and is manufactured from year 2001 to 2005 inclusive. drvexp(3) is the number of years of driving experience, 25 to 42 years.

The significant contributing factors for lowest risk are: male driver, driving experience of 25 to 42 years, and the vehicle is manufactured between the years 1991 and 2000 inclusive. The significant contributing factors for low risk are: lost control of vehicle, male driver and vehicle manufactured between the years 1991 and 2000 inclusive. The significant attributes for medium risk are: male driver, lost control of vehicle and driving a new car with reference to the duration of the data collected. The significant attributes for high risk are similar to medium risk level except for the additional attribute of wet road surface. Finally, the significant attributes for highest risk are: driving experience of 25 to 42 years, driving a new vehicle and hitting a tree.

Validation

A database of this magnitude and the use of free form text introduce issues with regards to the quality of the data and the associated results. Therefore, verification of results is required. The method selected to verify the rules was the accuracy measurement of the rules. The criteria for validation uses the statistical information collected during the analysis process such as accuracy and coverage. The classification power can be determined from the classification accuracy observed. The statistical information such as the accuracy and coverage are used to validate the rules. The accuracy is compared and is acceptable when the accuracy difference is within the defined threshold. The defined accuracy threshold is 80% with an allowance of ± 10.

The accuracy values for high and highest cost groups are zero as the ratio is too small and, thus, the nearest rounded number is zero. In addition, there are not many records with high and highest cost groups. This is not an ideal verification. Hence, a new table with 200 random selected rules with all the cost types is defined and verified with the rules obtained from the data analysis set.

The rules generated from the analysis data set are applied to the validation data to determine the classification accuracy. The new data set contains records for all cost groups and the accuracy obtained from the new data set has improved. The classification accuracy obtained is 90.4% with coverage of 57%. This is acceptable as the accuracy is within the threshold defined, 80% ± 10. This accuracy value implies that the risk assessment model will have an accuracy of 90.4% as it is defined based on the rules.
Discussion

The text mining process identified the contributing factors from the crash description of the insurance claim records. The factors from text mining are similar to the ones reported from road authorities such as Queensland Transport. Such similarity in contributing factors validates the accuracy of our text mining approach. Furthermore we identified new contributing factors which are not listed in Queensland Transport reports.

Rough set analysis produced a set of rules. The rules determine the dependency or relationship between contributing factors. The rules are selected based on the strength, thus, rules with strong strength are selected as it improves the prediction quality. The rules indicate the different combinations and relationships between contributing factors.

The rule for lowest crash risk suggests that the driver is a mature male, 50 to 59 years old, driving during evening peak hour and is involved in a collision with another vehicle. This may be a result of the high traffic volume during peak hour. The collision involved in this scenario is not a serious crash, which could be due to the fact that the driver is driving at a lower speed due to the higher traffic volume during peak hour. Thus, the impact of the crash is not as high as when a driver is driving at higher speed.

The lowest crash risk rule corresponds to a male driver, between 25 and 29 years old, driving between 12 pm and 4 pm, in a vehicle manufactured between 1991 and 2000. The most common type of crash is lost control of the vehicle. It seems probable that this type of crash could be due to inattention or the ‘afternoon lull’ which occurs between 1 pm and 5 pm. However it should be noted that even though the driver lost control of the vehicle, there was no collision with a vehicle or roadside objects.

The medium crash risk corresponds to drivers between 30 to 39 years old, driving between 7 pm to midnight, in a vehicle manufactured between 2001 and 2005 which is considered a new car with reference to the time the data is collected. The rule also indicates that the driver lost control of the vehicle and is intoxicated. Based on the time of crash and driver's intoxicated condition, the driver might be leaving from a party, dinner or company event where alcohol is served. A driver may misjudge the sharpness of the road curve and overestimates the speed, causing roll-over crashes or the vehicle to run-off road and hitting a kerb. This shows that alcohol and hitting a kerb or roadside object can increase the crash risk and the cost incurred.

The rule for highest crash risk indicates that driver between 25 and 29 years old, driving around midnight to 6 am and in a new car. Based on the time of the crash, the driver could be driving home after a party or from work. The driver could have been intoxicated and misjudged the sharpness of the curve while driving on a road curve, thus causing the vehicle to run off the road and hit a tree after a party. On the other hand, the driver may be overcome with fatigue and doze off, run off the road and hit a tree.

The rule predicting the collision type of the crash on a road curve suggests that the driver who drives during peak hour is more likely to be involved in a collision with another vehicle (not single vehicle crash). This could be attributed to the high traffic volume and fatigue.

The rules for hitting object type of crash describe a mature driver who is driving either at night or in the early hours of the morning and has a higher chance of being involved in hitting an object along the road, roadside or at the road. This could be caused by alcohol consumption or fatigue as indicated by the rule based on the time of crash.

The rules also indicate that most drivers lost control of the vehicle, and that factor led to the crash. This is a common contributing factor reported by road authorities’ reports; however, the relationships between contributing factors are not indicated. Thus, the relationships are identified with the rules obtained from rough set analysis. The rules suggest that loss of control could be due to speeding, wet or slippery road surface or avoiding animals.
The rules are verified with rough set theory and have an accuracy of approximately 90.4%.

**Conclusion and limitations**

This paper presents the application of text mining techniques to identify factors that contribute to crashes on road curves. The identified contributing factors are validated with a comparison with factors identified from a set of non-curve related crash records. This is to show that the factors identified are related to curve roads only. The contributing factors are: right-hand turn, narrow road, truck and avoiding kangaroo or dog.

The relationships between the contributing factors are identified with the rough set analysis process. The process produces a list of rules that represents the possible combination or relationships between the contributing factors. The rules are classified and represent a crash risk level. Each risk level is represented with a rule that has the highest support count. The selected rule indicates the combination of contributing factors that increases the crash risk.

A set of significant contributing factors is filtered from the set of rules. The significant contributing factors indicate the factors that have higher importance compared to the other factors in the data. The significant factors are identified for each risk level. The ones for the lowest risk are: male driver, driving experience of 25 to 42 years, and the vehicle is manufactured between the years 1991 and 2000 inclusive. As for the low risk are: lost control of vehicle, male driver and vehicle manufactured between the years 1991 and 2000 inclusive. As the risk increases, the ones for medium risk are: male driver, lost control of vehicle and driving a new car with reference to the duration of the data collected. The high risk has similar factors to the medium risk level except that one of the attributes is wet road surface. Lastly, the highest risk has driving experience of 25 to 42 years, driving a new vehicle and hitting a tree.

The set of rules are validated with accuracy measurement and the accuracy is approximately 90.4%. Based on the defined threshold of 80% with a ± 10, the rules are considered to be classified accurately.

One of the limitations of this study is the use of insurance claim records. These records could be biased towards the policy holder, who is hoping to reduce excess costs or preserve insurance rating. This could introduce bias into the results.

The number of records for crashes related to road curve is only less than 50% of the total number of crashes provided by the insurance company. Hence, the data available for analysis is limited.

**References**