Investigating the Pattern of Traffic Crashes under Rainy Weather by Association Rules in Data Mining

Subasish Das and Xiaoduan Sun
Department of Civil Engineering, University of Louisiana at Lafayette

Research Question

What are the key contributing factors for rainy weather crashes? What's the purpose of using association rules mining in place of conventional approaches?

Abstract

With a humid subtropical climate, the annual precipitation in Louisiana is about 64 inches, twice above the national average. Approximately 11 percent of total crashes in Louisiana happened during rainy weather, and nearly 25 percent of total fatal crashes happen in rainy weather annually. This paper demonstrates how to apply the association rules mining methods to discover hidden patterns in rainy weather crash data with eight years of Louisiana data (2004-2011). The findings of the study will help the highway authorities to determine countermeasure selection and safety improvement in adverse rainy weather.

Introduction

Rainy weather and wet roads are considered as hazardous conditions for driving. Due to the visual obstruction from rainfall and loss of surface friction, most vehicles slow down during rainfall but crashes still occur, which may be associated with the conditions. Measuring the added risk and identifying key crash contributing factors during showery weather has been challenging. The total reported number of traffic crashes in rainy weather in Louisiana is 11,398 in 2011 and 10,204 in 2010, respectively.

There are several ways to identify crash risk factors. The parametric models work well if the assumptions and model format are accurate to reflect the underlying relationships between dependent and independent variables. Violation of any assumption could lead to flawed or at least inadequate estimations. Specific nonparametric data mining techniques have been receiving increased attention from researchers in traffic safety because of no-predefined assumptions. One of the data mining techniques that has not been explored fully for crash data analysis is association rules. This method is concerned with the identification of interesting patterns from massive data.

Although number of studies have employed various data mining approaches in traffic safety research, generating significant rules with good visualization technique is missing. The National Traffic Safety Board (NTSB) reasoned that the risk of a fatal accident, nationwide, was about 3.9 to 4.5 times greater on wet pavement than on dry pavement. This study serves as a starting point to demonstrate the use of association rules mining to determine significant contributing rules that could present useful insight to the potential safety and traffic operation performance.

Association rules mining

Data mining is the process of identifying valid and understandable patterns in the data set. It helps in extracting and refining valuable knowledge from large data sets. Data mining involves machine learning, statistical knowledge, modeling concepts and database management. The methods can be classified into two main sections: descriptive and predictive. Association rules mining, a descriptive analytics, discovers significant rules showing variable category conditions that occur frequently together in a dataset. Many algorithms can be used to discover association rules from data to extract useful patterns. Apriori algorithm is one of the most widely used and famous techniques for finding association rules. Due to the explorative and eloquent nature, intelligible representation and visualization of the found patterns and models are essential for the successful mining process to make the results easy to understand. One important feature of the technique is that no variables are assigned as dependent or independent. The apriori algorithm for searching association rules is easy to interpret and the computations used are straightforward.

$$Rule: X \Rightarrow Y \quad Support = \frac{frq(X,Y)}{N} \quad Confidence = \frac{frq(X,Y)}{frq(X)} \quad Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Data Description

To identify important contributing factors for rainy weather crashes in Louisiana, a large dataset containing eight years of crash records (2004-2011) was obtained from the LADOTD.

- The final database contains 58,288 crash records with selected number of variables. The variable selection method used correlation matrix as a selection platform.
- To focus on the meaningful analysis, a set of key variables are selected such as the information on crash timing (day of the week), roadway characteristics (alignment, lighting), human factor (driver gender and age) and crash characteristics (crash severity and collision type).
- After a significant number of trials and errors, the minimum support for the rules was considered as 1 percent with the minimum confidence of 60 percent. One percent of minimum support means that no item or set of items will be considered frequent for the first analysis if it does not appear in at least 583 traffic crashes (1 percent of total 58,288 crash records).
- However a trial and error experiment indicates that setting minimum support too low will result in exponential growth of the number of items in the frequent item sets. By choosing different confidence values, a trial and error experiment showed that this parameter value gives rather stable results concerning the amount of rules generated by the algorithm. The purpose of post-processing the association rules set is to identify the subset of interesting rules in a generated set of noteworthy rules.

Major Findings

The exploration on the association rules mining might provide a better understanding of the risk factors. The results from this study is considered to be in use in future investigations with data mining researches in roadway safety.

Summary Results

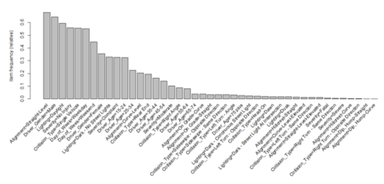

Figure 1: Item frequency plot.

Table 1: Summary chart of Association Rules Mining

$Support_{min}$=1% $Confidence_{min}$=60%				Lift ≥1								
				Support			Confidence			Lift		
Minlen	Maxlen	Rules (all)	Rules(Lift≥1)	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
1	1	2	2	0.644	0.661	0.678	0.644	0.661	0.678	1.000	1.000	1.000
2	2	81	59	0.010	0.109	0.412	0.600	0.715	0.849	1.002	1.147	1.482
3	3	423	330	0.010	0.054	0.257	0.601	0.722	0.922	1.001	1.176	1.660
4	4	871	692	0.010	0.033	0.152	0.600	0.721	0.939	1.000	1.190	1.691
5	5	767	622	0.010	0.022	0.092	0.601	0.723	0.939	1.001	1.207	1.693
6	6	217	179	0.010	0.015	0.036	0.600	0.720	0.928	1.000	1.217	1.934
7	7	6	6	0.011	0.011	0.012	0.631	0.740	0.843	1.072	1.249	1.471
All	All	2,367	1,890	0.010	0.034	0.678	0.600	0.722	0.939	1.000	1.194	1.934

Association Rules Visualization

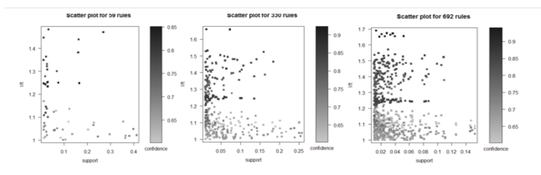

Figure 2: Scatter plot of the generated rules (a) 2-itemsets, (b) 3-itemsets, and (c) 4-itemsets.

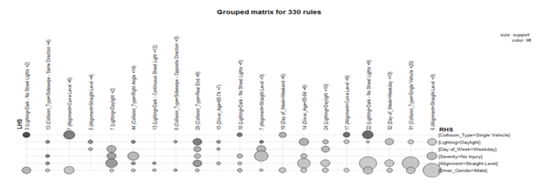

Figure 3: Grouped matrix for association rules for 3-itemsets.

Discussion

The association rules were generated in this study by using "arules" package in software R. The primary analysis demonstrates that the dataset has 58,288 rows with 43 items.

- Rules are created for two phases. When the minimum threshold of the lift is selected as one, the total counts of the rules are lowered.
- The most significant single variable category for the situation is found as single vehicle ROR crashes. This crash type is predominant for the rainy weather crashes in the presence of other roadway features such as on grade-curve aligned roadways, curved roadways, and roadways with no street lights at night. In rainy weather, PDO and sideswipe (same direction) crashes are significant in numbers. For drivers age 55 and above, most of the crashes during rainy weather happen in daylight. Moderate injuries are also dominant in single vehicle crashes. Roadways with poor illumination are associated with straight level aligned roadways for many rainy weather crashes. Young drivers (15-24) are vulnerable in ROR crashes when the roadways have poor illumination and are curve-aligned.
- The frequency of the rules generated for different itemsets and the statistics of support, confidence and lift are displayed in Figure 1 and Table 1, respectively.
- Inspecting all 1,890 rules manually is not a viable option. A straight-forward visualization of association rules involves a scatter plot (Figure 2) with two interest measures on the axes: the support values of the rules are on the x-axis and the lift values are on the y-axis.
- Figure 3 reveals the easier visualization of large sets of association rules from the grouped matrix. Balloon plots are drawn with antecedent groups as columns and consequents as rows (for example 3-itemsets). The color of the balloons represents the lift value and the size of the balloon shows the aggregated support.

Conclusion

This paper presents a preliminary investigation on how association rules mining techniques can be used to extract knowledge from the traffic crash data under a particular environmental condition. Some interesting findings are observed for crashes in rainy weather. Some of the findings verify the general perceptions on such types of crashes and a few findings are quite surprising. By observing the potential patterns in the discovered rules, the results can provide valuable insights into the underlying relationships between risk factors and crashes under particular conditions.

UNIVERSITY OF LOUISIANA Lafayette