Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool

Anurag Pande , Mohamed Abdel-Aty

Abstract

Data mining applications are becoming increasingly popular for many applications across a set of very divergent fields. Analysis of crash data is no exception. There are many data mining methodologies that have been applied to crash data in the recent past. However, one particular application conspicuously missing from the traffic safety literature until recently is association analysis or market basket analysis. The methodology is used by retailers all over the world to determine which items are purchased together. In this study, crashes are analyzed as supermarket transactions to detect interdependence among crash characteristics. The results from the analysis include simple rules that indicate which crash characteristics are associated with each other. The application is demonstrated using non-intersection crash data from the state of Florida for the year 2004. In the proposed methodology no variable needs to be assigned as dependent variable. Hence, it is useful in identifying previously unknown patterns in the data obtained from large jurisdictions (such as the State of Florida) as opposed to the data from a single roadway or intersection. Based on the association rules discovered from the analysis, it was concluded that there is a significant correlation between lack of illumination and high severity of crashes. Furthermore, it was found that under rainy conditions straight sections with vertical curves are particularly crash prone. Results are consistent with the understanding of crash characteristics and point to the potential of this methodology for the analysis of crash data collected by the state and federal agencies. The potential of this technique may be realized in the form of a decision support tool for the traffic safety administrators.

1. Introduction

The analysis of crash data using data mining techniques has been receiving increased attention from researchers (e.g., Abdel-Aty and Keller, 2005; Abdelwahab and Abdel-Aty, 2001; Chang and Chen, 2005; Chang, 2005). The usage of the data mining techniques has been largely limited to replace existing algorithms for classification problems (e.g., severity analysis conducted by Abdel-Aty and Keller, 2005; Abdelwahab and Abdel-Aty, 2001) and crash frequency estimation (Chang and Chen, 2005; Chang, 2005). The techniques used in most of these studies (e.g., neural network, classification tree) may be categorized, what Bayam et al. (2005) referred as predictive analysis, i.e., mapping a set of inputs to a specified output. On the other hand, descriptive analysis is used to discover groups of data objects (observations or variables) based on similarities/dissimilarities among these objects. Bayam et al. (2005) also discussed how neural network and classification trees (predictive data mining analysis) may be used for identifying crash patterns involving senior drivers. It was also pointed out that decision trees provide more understandable and explainable decisions compared to the neural networks. Hence, decision trees could be more useful for policy makers. Among the examples of descriptive data mining applications in traffic safety, Golob and Recker (2004) used clustering analysis for relating prevailing traffic conditions on freeways with type of collision most likely to occur.

One of the data mining techniques never utilized for crash data analysis until recently was the association analysis (Agrawal et al., 1993). It is part of the descriptive data mining analysis. The analysis involves looking into the data as transactions at the supermarket register to identify set(s) of items purchased together. The technique is also known as market basket analysis. In the proposed application, all the characteristics of crashes would be analyzed to search if certain characteristics tend to co-exist. In terms of understanding the results, association rules are preferred compared to cluster analysis because they provide specific and easy to describe relationships between crash attributes. One important feature of the technique is that no variables are assigned as dependent or independent.

The a priori algorithm for searching association rules is easy to understand and the computations used are straightforward. Due to explainable results and the ability to examine all potential relationships in the dataset this descriptive data mining may be a useful tool for policy makers. The application of the algorithm along with its potential future application as a decision support tool for policy makers is discussed in this paper. Crash data obtained from the Department of Highway Safety and Motor Vehicles (DHSMV), Florida are used in this study. The paper is organized as follows: In the next section benefits of this technique are discussed along with the objectives for which market basket analysis may be preferred over traditional techniques of crash data analysis. The methodology section describes the a priori rule discovery algorithm and the criteria for evaluating the discovered rules. The subsequent section is devoted to a detailed description of the data used in this study. The discovered association rules are presented in the ensuing section followed by the conclusions. The last section of the paper also discusses some future investigations that may help in fully exploiting the potential of this data mining methodology as a decision support tool in the area of traffic safety.

2. Motivation

"Good information properly used is one of the underpinnings of a sound traffic safety enterprise" (AASTHO strategic highway safety plan: goal 21). While there is a sufficient scope of improvement in the quality of the data being collected, all the state and the federal agencies do collect large amounts of crash data. Future policy initiatives are based on the conclusions drawn from these data. The data are often presented to the administrators in the form of multiple tables and illustrations to demonstrate latest trends in injuries and fatalities. In this study, we explore association rule mining for analyzing the data archived by one such agency (Department of Highway Safety and Motor Vehicles-FL) and discuss its potential as a decision support tool.

According to Hand et al. (2001), techniques such as association rule mining are better suited for analyzing observational data collected outside the purview of a designed experiment. Crash data from large jurisdictions (such as the State of Florida) are a good example of an observational database. Traditionally, studies dealing with crash data focus on establishing relationships between "dependent" and "independent" variables. However, the dichotomy used to categorize variables as dependent and independent variables is artificial and even arbitrary. Furthermore, it has been observed in the literature that correlations among independent variables significantly hamper the statistical analysis of crash data (e.g., Chang and Chen, 2005; Greibe, 2005). The correlations make it difficult to estimate the effect of different explanatory variables and may lead to incorrect conclusions (Greibe, 2005). Some recent studies have demonstrated that data mining techniques such as classification and regression tree (CART) can circumvent the problems arising from correlations (Chang and Chen, 2005; Chang, 2005). While the negative impact of correlations among independent variables can be countered using the aforementioned techniques, they provide no quantitative measure for these correlations.

Geurts et al. (2005) recently applied the association rule search algorithm to identify and differentiate between crash patterns in and outside of the "black" zones. The analysis was based on 1861 injury crashes that occurred in a small province, south of Brussels (Belgium). However, their analysis of the crash data was too similar to traditional marketing applications. In other words, thresholds applied on the rule evaluation criteria (described in the next section) were closer to the ones used for marketing applications (also see the section titled analysis). In this study, we apply this algorithm on the crash data by clearly differentiating it from the traditional marketing applications.

As observed later from the results, the proposed methodology (i.e., association rules mining) can potentially identify relationships that are not well known from the traffic safety literature. Without restricting the nature of variables (as dependent or independent) one can find valuable relationships which would otherwise remain elusive. The market basket analysis also results in rules that are easy to understand. Despite the advantages, association rule mining is NOT intended to be a replacement for other techniques used for statistical analyzes of crash data. Instead, it is an efficient tool for analyzing huge database of crash characteristics from jurisdictions such as a state DOT.

3. Methodology

Association discovery is the identification of sets of items that occur together in a given event or record. This technique is also known as market basket analysis. On-line transaction processing systems at the supermarkets often provide the data sources for association discovery. Association rules are based on the relative frequency of the number of times the sets of items occur alone and in combination in a database. They are expressed as follows: "if item A is part of an event then item B is also a part of

the event X% of the time". We would represent the aforementioned rule as "A → B", where A is the antecedent on the LHS and B is the consequent on the RHS. Note that one can have multiple items, i.e., a set of items, as antecedent and consequent in a rule. For further clarification, here are some hypothetical examples of the association rules:

- If a customer buys beer, then he/she also buys chips ("beer → chips").
- A grocery chain may find that 80% of all shoppers will buy a jar of salsa when they purchase a bag of tortilla chips ("bag of tortilla chips → jar of salsa").

It is worth mentioning that these rules should not be interpreted as a direct causation, but as associations between the sets of items (SAS Institute, 2001).

Agrawal et al. (1993) first introduced the framework to search for association rules in large databases based on the a priori algorithm. A priori algorithm uses simple and step-by-step ways to repetitively examine candidate item-sets to find frequent item-sets. Then, it uses the new candidate item-sets produced using frequent item-sets to find new frequent item-sets until no more new item-sets can be produced. The concepts of support and confidence are central to association rules. Support is a measure of how frequently any given combination of antecedent and consequent occurs in a database. Confidence is defined by the percentage of cases in which a consequent appears given that the antecedent has occurred. It essentially measures the strength of an association rule. The framework proposed by Agrawal et al. (1993) consisted of only these two parameters for evaluation of the rules generated by the algorithm. However, Brin et al. (1998) introduced a third evaluation parameter, which was referred to as "interest" or "lift". These three criteria are defined with an example as follows:

Suppose a hypothetical crash database consist of 100,000 crashes and out of these crashes 20,000 of them occurred under "dark without street lights" and 5000 of them were fatal. Out of these 5000 fatal crashes, 2000 occurred under dark without street lights. Now consider the rule "dark without street lights → fatal crashes" for this database. In this rule, "dark without street lights" is the antecedent while the "fatal crashes" is the consequent. The support for the rule is defined as the percentage of all crashes that were both fatal and occurred on a dark street without lights. For the aforementioned hypothetical rule, support would be 2% (2000/100,000 = 0.02). Confidence for the rule is defined as the percentage of fatal crashes among all crashes that occurred under dark conditions on roadways with no street lights. The number of such crashes is 2000 and hence in this database, the confidence for the aforementioned rule would be 10% (2000/20,000 = 0.10). As we shall note later, the most important criterion of the three in the context of the present problem (i.e., crash data analysis) is "interest" or "lift". The later of the two terms is used from here on. The lift of the rule measures the statistical dependence of the rule by relating the observed frequency of co-occurrence to the expected frequency of the co-occurrence under the assumption of conditional independence. The higher the lift for a rule, the more interesting the rule would be since it would indicate how more often the two characteristics are part of the same crash than if these events were statistically independent. In mathematical terms the lift is defined as follows:

$$\text{Lift} = \frac{P(\text{fatal crash}|\text{dark with no street lights})}{P(\text{fatal crash})} = \frac{P(\text{fatal crash} \cap \text{dark with no street lights})}{P(\text{fatal crash}) * P(\text{dark with no street lights})}$$

The definitions of the three parameters, based on this example, are further clarified in Fig. 1. The association rule discovery is the process of finding strong associations with a minimum support and/or confidence and lift value greater than one. Minimum support controls the number of observations that must contain the antecedent and consequent combination; while minimum confidence controls the predictive power of the rules.

It is desirable for the rules to have a large confidence factor, a high level of support, and a lift value greater than one. Since some events of interest in traffic safety analysis are very rare (e.g., "crashes with fatal injury"); the support for some rules of interest could be quite low. It essentially means that the lift value is more important for determining the strength of an association rule than the other two criteria. Hence, in the present application the rules should be evaluated based on the 'lift' values. It is not to say that the other two criteria are of no importance. The rules 'discovered' by the algorithm still need to have support greater than a minimum threshold. The threshold, however, would be set much lower (close to 1%) compared to a marketing application. The threshold ensures that the pattern identified by a rule is observed in the database with at least some reasonable frequency. If one only relies on the lift value and not use a threshold for minimum support it is possible to identify rules that are based on very few crashes. These rules would be of little practical value.

Support considers only the combination of crash characteristics and not the direction. In other words, two rules with flipped antecedent and consequent will both have the same support. The confidence is useful in differentiating between such rules. Consider a customer database with 25% support for the combination of two products say, beer and lime. This could mean that 25% of all customers buy both beer and lime, and no one buys beer without buying lime. In that case, it would be a good rule. But what if 100% of customers buy beer and only 25% of those buy lime? In which case, it would not be a good rule, even though the support is still 25%. The fact that a customer bought beer does not really reveal whether they will buy lime. The parameter confidence provides a measure for how confident one can be of the fact that given a customer has purchased one product, they will also purchase the

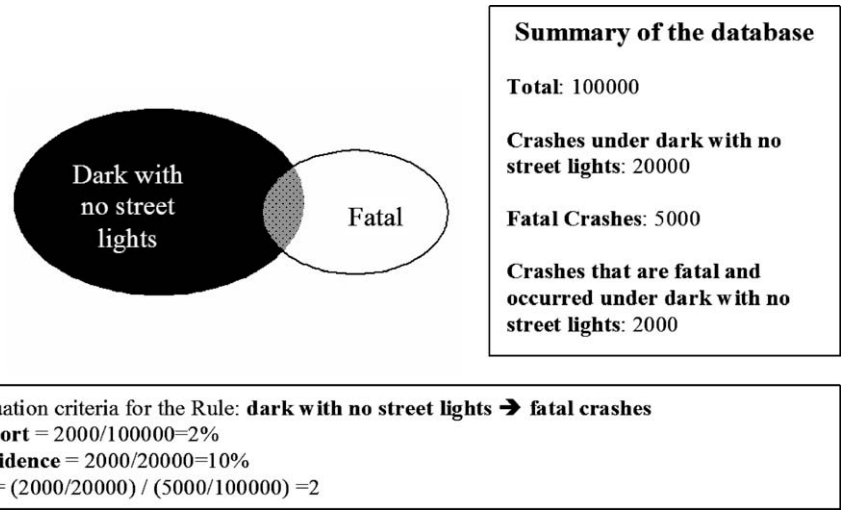

Fig. 1. Three association rule evaluation criteria based on a hypothetical example crash database.

other product. Confidence is especially important when dealing with characteristics that exist in a large proportion of crashes such as "clear weather" (68.72%).

It is worth emphasizing again that the objective is not to establish any specific relationship(s) but to examine how various crash characteristics are associated with each other. It is in contrast with traditional multivariate modeling of the crash data where a relationship (between the so-called dependent and independent variables) is sought and model parameters are estimated to specify the relationship. In association discovery the goal is to get some information – any real information – out of the data. In the following section, details of the crash data used in this study are described.

4. Data description

The data used herein are obtained from DHSMV that maintains a database for all crashes reported in the state of Florida. In this study, we are using database for crashes that did not occur on intersections or ramps. The database included crashes with five severity levels, varying from "no injury" to crashes involving "fatal injury". While there might be some reporting bias leading to under-reporting of the least severe (i.e., No injury) crashes (Abdel-Aty and Keller, 2005), the frequency of the crashes belonging to the other severity levels is expected to be fairly accurate. Since relative frequency of crash characteristics is one of the critical aspects of the association rule discovery process, "no injury" crashes were removed from the database. Due to accurate reporting of the remaining crashes the database may now be expected to have correct proportions of crashes belonging to remaining four severity levels. The part of the database contains following characteristics about each crash:

- Crash injury severity (possible injury, non-incapacitating injury, incapacitating injury, and fatal injury).
- Light conditions (daylight, dusk, dawn, dark with street light, and dark with no street light).
- Weather conditions (clear, cloudy, rain, and fog).
- Traffic-way character (straight level, straight grade, curve level, and curve grade).
- Separation of traffic (divided and undivided highway).

In traditional multivariate modeling of the crash data, "crash injury severity" would have been the dependent variable with the other four being the independent variables. In this study, however, every category within a nominal variable would be treated as a 'product' in a 'market basket'. A crash that occurred daylight, clear weather, straight level, divided highway and involved possible injury, can be treated as a transaction during which these 'products' were purchased.

Table 1 summarizes the information available for 59,679 non-intersection crashes. To find interesting patterns one would look for the crash characteristics that occur together significantly more often than they would if they had been statistically independent of each other. Therefore, the market basket analysis may be understood as a more sophisticated and efficient substitute for contingency tables. A contingency table, also called a cross-reference table, is a table showing the number of records for each level combination of two or more categorical variables that constitute the table. As the size of the contingency table grows, it becomes difficult to keep track of the results. Association discovery may be seen as a process of looking through all possible multi-way contingency tables and filtering out the most 'interesting' of the conclusions.

It should be re-emphasized that even though support for some association rules in a crash database might be low due to rare occurrence of the characteristics included in the rule, they might be of significant interest (e.g., rules involving "fatal injury"). Also, note that sometimes the 'discovered' rules might be obvious and hence useless (e.g., "mother → female"). Therefore, once the association

Table 1
Summary of crash characteristics

Lighting condition		Frequency	Percent	Cumulative frequency	Cumulative percent
01	Daylight	38,859	65.11	38,859	65.11
02	Dusk	1574	2.64	40,433	67.75
03	Dawn	882	1.48	41,315	69.23
04	Dark (street light)	11,123	18.64	52,438	87.87
05	Dark (no light)	7241	12.13	59,679	100
Weather					
01	Clear	41,009	68.72	41,009	68.72
02	Cloudy	11,999	20.11	53,008	88.82
03	Rain	6074	10.18	59,082	99.00
04	Fog	324	0.54	59,406	99.54
05	All other	273	0.46	59,679	100
Traffic-way character					
01	Straight-level	48,749	81.69	48,749	81.69
02	Straight-upgrade/downgrade	5057	8.47	53,806	90.16
03	Curve-level	4509	7.56	58,315	97.71
04	Curve-upgrade/downgrade	1364	2.29	59,679	100
Divided/undivided highway					
1	Divided highway	34,150	57.22	34,150	57.22
2	Undivided highway	25,529	42.78	59,679	100
Crash injury severity					
2	Possible injury	27,402	45.92	27,402	45.92
3	Non-incapacitating evident injury	20,691	34.67	48,093	80.59
4	Incapacitating injury	9838	16.48	57,931	97.07
5	Fatal injury	1748	2.93	59,679	100

rules have been discovered using the a priori algorithm they need to be vetted or "post-processed" carefully for valuable information.

5. Data preparation

To perform association discovery using the SAS enterprise miner the input data set must have a separate observation for each product purchased by each customer (SAS Institute, 2001). Correspondingly, the format of the original data obtained form the DHSMV database had to be changed. As mentioned earlier, five categorical variables are included in the analysis. Lighting conditions have five levels, weather conditions, traffic-way character, and crash injury severity and have four levels each, and separation of traffic has two levels. Hence, the database of 59,679 crashes with five variables will be expanded to have 298,395 (=59,679*5) observations. The format of the raw dataset (with a sample of two crashes) and the one prepared for the association rules search is provided in Table 2a and Table 2b, respectively. The recoded data are subjected to the a priori algorithm to search for association rules. It is worth mentioning that each crash is being treated as a transaction while the corresponding categories of the five variables are treated as 'product' purchased during that transaction.

Table 2a
Sample of crash data

Crash no.	Lighting condition	Weather	Traffic-way character	Divided/ undivided highway	Crash injury severity
1	1	1	1	1	2
2	1	1	1	1	4

Table 2b
Sample of crash data recoded for association analysis

No.	Product (condition)	Transaction Id (crash no.)	Variable category
1	Daylight	1	1 Lighting condition
2	clear_weather	1	2 Weather
3	Straight_Level	1	3 Traffic-way character
4	divided_HW	1	4 Divided/undivided highway
5	Possible_injury	1	5 Crash injury severity
6	Daylight	2	1 Lighting condition
7	clear_weather	2	2 Weather
8	Straight_Level	2	3 Traffic-way character
9	divided_HW	2	4 Divided/undivided highway
10	Incap_injury	2	5 Crash injury severity

6. Analysis

Prior to searching for the rules, minimum thresholds for support and confidence were specified. The threshold values used in the analysis are 0.90% and 10%, respectively. It means that no rules with support <0.90% and/or confidence <10% would be considered irrespective of their lift

values. The role of these thresholds was discussed with the concepts of support and confidence in the "methodology" section. These thresholds are lower than the values typically used in market basket analysis due to our interest in rare crash characteristics (such as a fatal injury). It is worth mentioning that Geurts et al. (2005) used 5% as the threshold on support parameter for their analysis of crashes in and outside the "black" zones. The 5% threshold is closer to the value used in marketing applications. This is one reason why they were not able to 'discover' any patterns related to driver, passenger, and/or victim fatality.

Another specification used in the SAS enterprise miner (SAS Institute, 2001) is that the upper limit on the 'products' included in a single rule. The upper limit was set at four and therefore, 2-product, 3-product and 4-product rules would be identified. In the next section, the rules uncovered from the dataset described in the previous section are presented. The rules are represented in the following form: "antecedent → consequent ($L = x$, $S = y$, $C = z$)", where x, y, and z represent the values of lift, support and confidence for the corresponding rule. It is worth mentioning that antecedent and the consequent in the rules could be a single 'product' (i.e., the category of a variable such as "fatal crash") or a set of 'products' (such as "fatal crash and dark with no street lights and level grade"). The rules discovered from this dataset based on the a priori algorithm are shown in Tables 3–5. The tables include the following parameters:

- Lift.
- Support (%).
- Confidence (%).
- Transaction count: number of transaction in which the particular combination of 'products' occur.
- Rule: antecedent → consequent.

The rules are sorted by the descending lift values. Tables 3–5 show "2-product", "3-product" and "4-product" rules, respectively. It is worth mentioning that more rules were 'discovered' by the algorithm than the ones shown in Tables 3–5. Note that the rules with lift values close to 1.0 are of little interest. Only rules with lift greater than or equal to 1.25 are shown in Table 3. The next highest lift for any 2-product rule was only 1.16. Due to this significant drop in the lift value only 15 rules are included in Table 3. Similar procedure was used to determine how many 3-product and 4-product rules to include in Tables 4 and 5. Some of the remarkable rules shown in the Tables are discussed in the following section.

7. Discussion of the rules discovered

The first rule in Table 3 indicates that if a crash results in fatal injury it is more likely to have occurred on dark with no street light ("fatal injury → dark with no street light ($L = 2.73$, $S = 0.97$, $C = 33.12$)"). In this regard, two other rules are worth mentioning "incapacitating injury → dark with no street light ($L = 1.53$, $S = 3.06$, $C = 18.54$)" "non-incapacitating injury → dark with no street light ($L = 1.04$, $S = 4.37$, $C = 12.61$)". Note that the later of the two rules is not included in Table 3 due to its low lift value. Based on these three rules and the corresponding lift values, which decrease with the antecedent severity levels, it may be inferred that under dark conditions with no street lights crashes are likely to be more severe. It also indicates that installing street lights could help in reducing the severity of crashes.

Other interesting "2-product" rules include "curve level → dark with no street light ($L = 2.09$, $S = 1.92$, $C = 25.35$)", "curve level → undivided highway ($L = 1.49$, $S = 4.80$, $C = 63.58$)", and "curve level → incapacitating injury ($L = 1.47$, $S = 1.84$, $C = 24.31$)". These rules indicate that the crashes that occur on section with level grade and horizontal curve are more likely to occur on undivided highway, under dark with no street lights, and incur incapacitating injury. Another interesting rule indicates that if a crash occurred during rain it is likely to occur on straight roadway sections with vertical curve (Rule #10; Table 3).

Table 3
List of "2-product" rules

Rule #	Lift	Support (%)	Confidence (%)	Transaction count	Rule
1	2.73	0.97	33.12	579	fatal_injury → DarkNoSL
2	2.09	1.92	25.35	1143	Curve_Level → DarkNoSL
3	2.09	1.92	15.79	1143	DarkNoSL → Curve_Level
4	1.53	3.06	25.19	1824	DarkNoSL → Incap_injury
5	1.53	3.06	18.54	1824	Incap_injury → DarkNoSL
6	1.49	4.8	11.23	2867	undivided_HW → Curve_Level
7	1.49	4.8	63.58	2867	Curve_Level → undivided_HW
8	1.47	1.84	24.31	1096	Curve_Level → Incap_injury
9	1.47	1.84	11.14	1096	Incap_injury → Curve_Level
10	1.4	1.2	11.82	718	rain → Straight_grade
11	1.4	1.2	14.2	718	Straight_grade → rain
12	1.38	7.17	16.76	4278	undivided_HW → DarkNoSL
13	1.38	7.17	59.08	4278	DarkNoSL → undivided_HW
14	1.25	2.13	25.09	1269	Straight_grade → cloudy
15	1.25	2.13	10.58	1269	cloudy → Straight_grade

Table 4
List of "3-product" rules

Rule #	Lift	Support (%)	Confidence (%)	Count	Rule
1	2.77	1.5	19.85	895	Curve_Level → undivided_HW & DarkNoSL
2	2.77	1.5	20.92	895	undivided_HW & DarkNoSL → Curve_Level
3	2.57	1.5	12.36	895	DarkNoSL → undivided_HW & Curve_Level
4	2.57	1.5	31.22	895	undivided_HW & Curve_Level → DarkNoSL
5	2.18	1.33	26.43	792	clear_weather & Curve_Level → DarkNoSL
6	2.18	1.33	10.94	792	DarkNoSL → clear_weather & Curve_Level
7	2.11	1.33	17.56	792	Curve_Level → clear_weather & DarkNoSL
8	2.11	1.33	15.94	792	clear_weather & DarkNoSL → Curve_Level
9	2.09	1.26	15.77	754	undivided_HW & Incap_injury → Curve_Level
10	2.09	1.26	16.72	754	Curve_Level → undivided_HW & Incap_injury
11	2.01	1.96	16.13	1168	DarkNoSL → undivided_HW & Incap_injury
12	2.01	1.96	24.44	1168	undivided_HW & Incap_injury → DarkNoSL
13	1.83	1.5	78.3	895	DarkNoSL & Curve_Level → undivided_HW
14	1.66	1.96	27.3	1168	undivided_HW & DarkNoSL → Incap_injury
15	1.66	1.96	11.87	1168	Incap_injury → undivided_HW & DarkNoSL
16	1.62	0.91	10.72	542	Straight_grade → rain & divided_HW
17	1.62	0.91	13.72	542	rain & divided_HW → Straight_grade
18	1.61	0.93	12.26	553	Curve_Level → undivided_HW & cloudy
19	1.61	0.93	12.19	553	undivided_HW & cloudy → Curve_Level
20	1.61	1.26	68.8	754	Incap_injury & Curve_Level → undivided_HW
21	1.6	1.26	26.3	754	undivided_HW & Curve_Level → Incap_injury
22	1.58	0.91	16.12	542	divided_HW & Straight_grade → rain
23	1.57	2.16	25.96	1290	clear_weather & DarkNoSL → Incap_injury
24	1.57	2.16	13.11	1290	Incap_injury → clear_weather & DarkNoSL
25	1.56	1.84	24.42	1101	Curve_Level → undivided_HW & Non_Incap_injury
26	1.56	1.84	11.78	1101	undivided_HW & Non_Incap_injury → Curve_Level
27	1.56	2.16	18.91	1290	clear_weather & Incap_injury → DarkNoSL
28	1.56	2.16	17.82	1290	DarkNoSL → clear_weather & Incap_injury
29	1.55	3.32	66.1	1981	clear_weather & Curve_Level → undivided_HW
30	1.53	1.26	25.16	754	clear_weather & Curve_Level → Incap_injury
31	1.5	2.21	24.78	1316	Straight_Level & DarkNoSL → Incap_injury
32	1.5	2.21	13.38	1316	Incap_injury → Straight_Level & DarkNoSL
33	1.5	1.96	64.04	1168	Incap_injury & DarkNoSL → undivided_HW

A similar rule is discovered for the cloudy weather, although the lift value for the rule is only 1.25 (Rule #15; Table 3). The analysis also uncovers that undivided roadways in general can be expected to have crashes under dark conditions without street lights (Rule #12; Table 3). The reason for this association could be that the undivided roads are more likely to be without street lights thereby increasing the exposure for such conditions.

In Table 4, the rule with highest lift value is "curve level → undivided highway and dark with no street light ($L = 2.77$, $S = 1.5$, $C = 19.85$)" indicates that a crash on level grade and horizontal curve is more likely to be on undivided highway as well as under dark with no street lights. It once again indicates that it might be worth considering dividing the highways and installing lights on the level sections with horizontal curves. Note that in a marketing strategy this rule would not be given much attention due to its low support value. Low level of support means that the constituent of the rules is rare. This highlights the difference between association rules discovery in crash data analysis and marketing application due to emphasis of the former on the rare harmful events.

Another interesting set of rules is related to "rain and divided highway → straight grade ($L = 1.62$, $S = 0.91$, $C = 13.72$)" The straight grade means that the roadways only have a vertical curve and no horizontal curve. The rule implies that crash on a divided highway during rain is more likely to have occurred on a vertical curve. Rule #10 (Table 4) suggests that a crash on a level road with a horizontal curve is more likely to occur on undivided highway and involve incapacitating injury. It may be observed from Table 5 that the rules with four variables provide and further substantiate aforementioned conclusions. A closer look at the list of rules would also indicate that the many rules are 'repeated' with flipped antecedent and consequent (i.e., the LHS of the rules becomes the RHS). The only evaluation criteria that changes between two such rules would be the confidence for the rule. It re-emphasizes that these rules should not be interpreted as the 'causality' but as associations. Inferences regarding 'causality' require domain knowledge from traffic safety analysts and highway design engineers. Based on the rules discovered, the following 'actionable' conclusions may be drawn:

- During rainy/cloudy conditions the roadways with a vertical curve are particularly crash prone.
- Dark conditions without street lights are prone to more severe crashes.

Table 5
List of "4-product" rules

Rule #	Lift	Support (%)	Confidence (%)	Count	Rule
1	3	1.08	15.05	644	undivided_HW & DarkNoSL → clear_weather & Curve_Level
2	3	1.08	21.49	644	clear_weather & Curve_Level → undivided_HW & DarkNoSL
3	2.8	1.08	14.28	644	Curve_Level → undivided_HW & clear_weather & DarkNoSL
4	2.8	1.08	21.18	644	undivided_HW & clear_weather & DarkNoSL → Curve_Level
5	2.7	1.08	12.96	644	clear_weather & DarkNoSL → undivided_HW & Curve_Level
6	2.7	1.08	22.46	644	undivided_HW & Curve_Level → clear_weather & DarkNoSL
7	2.68	1.08	32.51	644	undivided_HW & clear_weather & Curve_Level → DarkNoSL
8	2.19	0.88	11	526	undivided_HW & Incap_injury → clear_weather & Curve_Level
9	2.19	0.88	17.55	526	clear_weather & Curve_Level → undivided_HW & Incap_injury
10	2.1	1.4	16.78	834	clear_weather & DarkNoSL → undivided_HW & Incap_injury
11	2.1	1.4	17.45	834	undivided_HW & Incap_injury → clear_weather & DarkNoSL
12	2.03	0.88	11.67	526	Curve_Level → undivided_HW & clear_weather & Incap_injury
13	2.03	0.88	15.3	526	undivided_HW & clear_weather & Incap_injury → Curve_Level
14	2	1.4	24.27	834	undivided_HW & clear_weather & Incap_injury → DarkNoSL
15	2	1.4	11.52	834	DarkNoSL → undivided_HW & clear_weather & Incap_injury
16	1.9	1.08	81.31	644	clear_weather & DarkNoSL & Curve_Level → undivided_HW
17	1.85	1.32	16.49	788	undivided_HW & Incap_injury → Straight_Level & DarkNoSL
18	1.85	1.32	14.84	788	Straight_Level & DarkNoSL → undivided_HW & Incap_injury
19	1.84	1.32	10.88	788	DarkNoSL → undivided_HW & Incap_injury & Straight_Level
20	1.84	1.32	22.32	788	undivided_HW & Incap_injury & Straight_Level → DarkNoSL
21	1.82	1.08	56.34	644	DarkNoSL & Curve_Level → undivided_HW & clear_weather
22	1.71	1.4	12.23	834	clear_weather & Incap_injury → undivided_HW & DarkNoSL
23	1.71	1.4	19.5	834	undivided_HW & DarkNoSL → clear_weather & Incap_injury
24	1.66	1.4	27.43	834	undivided_HW & clear_weather & DarkNoSL → Incap_injury
25	1.64	1.32	26.98	788	undivided_HW & Straight_Level & DarkNoSL → Incap_injury
26	1.63	0.88	69.76	526	clear_weather & Incap_injury & Curve_Level → undivided_HW
27	1.63	1.28	25.46	763	clear_weather & Curve_Level → undivided_HW & Non_Incap_injury
28	1.61	0.88	26.55	526	undivided_HW & clear_weather & Curve_Level → Incap_injury
29	1.6	0.88	18.35	526	undivided_HW & Curve_Level → clear_weather & Incap_injury
30	1.56	0.94	24.36	561	Daylight & Curve_Level → undivided_HW & Non_Incap_injury
31	1.55	1.58	17.72	941	Straight_Level & DarkNoSL → clear_weather & Incap_injury
32	1.55	1.58	13.79	941	clear_weather & Incap_injury → Straight_Level & DarkNoSL
33	1.55	1.58	25.54	941	clear_weather & Straight_Level & DarkNoSL → Incap_injury
34	1.55	0.88	47.99	526	Incap_injury & Curve_Level → undivided_HW & clear_weather
35	1.52	1.28	65.1	763	clear_weather & Non_Incap_injury & Curve_Level → undivided_HW
36	1.51	1.4	64.65	834	clear_weather & Incap_injury & DarkNoSL → undivided_HW
37	1.51	0.95	64.46	564	clear_weather & Possible_injury & Curve_Level → undivided_HW

- Sections with horizontal curve are prone to crashes involving incapacitating injury.

It is reasonable to compare some of the conclusions from this analysis to the findings of the past studies. As mentioned earlier, most of the studies dealing with severity of crashes have used severity as the dependent variable. For example, Shankar et al. (1996) concluded that night time conditions with no street lights increase the probability of property damage (i.e., No injury) crashes on a 61 km study section of I-90 in the state of Washington. It was argued that since the most dangerous portion of this major freeway was likely to be illuminated, a positive correlation between the absence of illumination and the likelihood of a property damage, only crash, is increased. Also, one of the previous studies by Abdel-Aty (2003) based on crash data from roadway sections in the Central Florida area did not find the variable "weather" as significantly affecting the severity levels. Other results from the study about the relationship of the severity levels with horizontal curve and lighting conditions were consistent with the results obtained here. These references highlight the potential of market basket analysis to uncover patterns, whether related to severity or crash frequency, which may sometimes remain undetected by the traditional approach to crash data analysis (i.e., with pre-specified input and output variables). It is also worth mentioning that the differences between lift values for the rules involving different antecedent (for example, night time and day time) but same consequent (e.g., fatal crash) could provide a measure akin to "measure of effect" estimated through multivariate statistical models.

There are other methods such as "chi-square test of count tables" for examining the presence of associations among variables. However, some issues need to be considered when using the chi-square test. A significant overall chi-square test would indicate that the categorical variables forming the contingency table are not independent, but provides no information as to whether the lack of independence occurs throughout the table or only in a specific part. In the proposed algorithm, each category of a nominal variable is treated as a different 'product'. Hence, the

discovered rules would provide us, for example, not only if weather and vertical alignments are correlated but also if crashes under rain (a category within the variable weather) are more likely to also occur on downgrade (a category within the variable vertical alignment).

Since the market basket analysis has never been applied before to crash data from the United States, further explorations with the data could reveal more patterns of interest. Further 'mining' could be carried out with more parameters to build on the promising results obtained here. These parameters could include at-fault driver age-groups, not-at-fault driver age-groups, gender, etc. The analysis may be extended to intersection crashes as well. Also, in this study we have only focused on association rules with lift values greater than unity. The algorithm may be modified to also provide rules with lift $<< 1$. In marketing applications, such rules are aplenty but generally useless (since they indicate the products that do not sell together). For crash data analysis, characteristics that generally do not occur together would also be of interest.

8. Conclusive remarks and future scope

In this study, the application of market basket analysis on crash data is demonstrated using the data from the State of Florida. The a priori algorithm to search for association rules in the crash data is applied in this study with crucial changes made to thresholds used on rule evaluation criteria. These modifications were directed towards making association rule mining more suitable for crash data analysis. More specifically, since the dataset involves rare events of interest such as fatal crashes, the thresholds used for minimum support were lowered. The lowering of threshold allows the analyst to be able to discover association involving such rare events.

The application of market basket analysis could be very useful in detecting patterns in the crash data obtained from a large jurisdiction. Since these data are already being collected by various agencies around the world, association discovery analysis becomes all the more suitable. It enables one to look at the data without any 'prejudice' and without limiting the amount of information data could potentially provide. Retailers around the world have found this to be a good tool to estimate the items that are purchased together. Similarly, association rules can be useful for agencies looking into crash patterns to identify policy initiatives for reducing frequency and severity of crashes. The a priori algorithm is indeed a systematic way of exhaustively examining cells of all possible contingency tables (similar to the ones found in publications similar to Traffic Safety Facts, 2001) within a large dataset. The only difference is that the results are presented in the form of association rules. Hence, it is a simple but efficient approach to search for patterns in the statewide/nationwide crash data.

In this study, the application was demonstrated using five different variables in the crash database. With five variables the number of discovered rules was manageable. However, to fulfill the potential of association rule mining as a traffic safety decision support tool; the algorithm, of course, would have to be applied to datasets with many more variables. The lower minimum support threshold (anywhere close to the one used in this study) would result in significantly higher number of discovered rules. Large number of rules would require more sophisticated ways to mine the patterns within the discovered association rules. In this regard, additional measures of finding 'interesting' rules in the database would need to be explored. These measures include Gain measure and Conviction proposed by Fukuda et al. (1996) and Brin et al. (1997), respectively. More such measures have also been documented by Bayardo and Agrawal (1999). In this regard, some studies in the field of bioinformatics have also proposed detailed algorithms for comprehensive post-processing of discovered rules. These algorithms include the ones proposed by Tuzhilin and Adomavicius (2002) and Tuzhilin and Liu (2002). Review of these studies reveals that algorithms for post-processing of rules along with the additional measures of interestingness are highly context dependent. Developing such analyzes in the context of crash data is not a trivial matter and is, therefore, worthy of future investigations.

These investigations are necessary for developing decision support tools based on association rule mining. As with the market basket analysis in the retail sector where it is up to the data owners to re-shelve their items based on the results, it would be up to the agencies to act on these broad patterns discovered from the data to develop policy initiatives and/or specific solutions for reduction in injuries and fatalities on roadways.

References

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34 (5), 597–603.

Abdel-Aty, M., Keller, J., 2005. Exploring the overall and specific crash severity levels at signalized intersections. Accident Analysis and Prevention 37 (3), 417–425.

Abdelwahab, H., Abdel-Aty, M., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record 1746, 6–13.

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD, Washington DC, pp. 207–216.

Bayam, E., Liebowitz, J., Agresti, W., 2005. Older drivers and accidents: a meta analysis and data mining application on traffic accident data. Expert Systems with Applications 29 (3), 598–629.

Bayardo Jr., R., Agrawal, R., 1999. Mining the most interesting rules. In: Proceedings of the 1999 ACM-SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, pp. 145–154.

Brin, S., Motwani, R., Ullman, J., Tsur, S., 1997. Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the 1997 ACM-SIGMOD International Conference on the Management of Data, pp. 255–264.

Brin, S., Motwani, R., Silverstein, C., 1998. Beyond market baskets: generalizing association rules to dependence rules. Data Mining and Knowledge Discovery 2 (1), 39–68.

Chang, Li-Y., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Safety Science 43 (8), 541–557.

Chang, Li-Y., Chen, Wen-C., 2005. Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research 36 (4), 365–375.

Fukuda, T., Morimoto, Y., Morishita, S., Tokuyama, T., 1996. Data mining using two-dimensional optimized association rules: scheme, algorithms, and visualization. In: Proceedings of the 1996 ACM-SIGMOD International Conference on the Management of Data, pp. 13–23.

Geurts, K., Thomas, I., Wets, G., 2005. Understanding spatial concentrations of road accidents using frequent item sets. Accident Analysis and Prevention 37 (4), 787–799.

Golob, T., Recker, W., 2004. A Method for relating type of cash to traffic flow characteristics on urban freeways. Transportation Research – Part A, Policy and Practice 38, 52–80.

Greibe, P., 2005. Accident prediction models for urban roads. Accident Analysis and Prevention 35 (2), 273–285.

Hand, D., Mannila, H., Smyth, P., 2001. Principles of Data Mining. The MIT Press, Cambridge, MA.

SAS Institute, 2001. Getting started with enterprise miner software. In: Release 4.1. SAS Institute, Cary, NC.

Shankar, V., Mannering, F., Barfield, W., 1996. Statistical analysis of accident severity on rural freeways. Accident Analysis and Prevention 28 (3), 391–401.

Traffic Safety Facts, 2001. National Highway Traffic Safety Administration. National Center for Statistics and Analysis.

Tuzhilin, A., Adomavicius, G., 2002. Handling very large numbers of association rules in the analysis of microarray data. In: Proceedings of the 2002 ACM-SIGKDD International Conference on Knowledge discovery and Data Mining, pp. 396-404.

Tuzhilin, A., Liu, B., 2002. Querying multiple sets of discovered rules. In: Proceedings of the 2002 ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 52–60.