Application of data mining techniques for real-time crash risk assessment on freeways

A. Pande and M. A. Abdel-Aty

Abstract

Data mining is the analysis of large "observational" datasets to find unsuspected relationships that might be useful to the data owner. It typically involves analysis where objectives of the mining exercise have no bearing on the data collection strategy. Freeway traffic surveillance data collected through underground loop detectors is one such "observational" database maintained for various ITS (Intelligent Transportation Systems) applications such as travel time prediction etc. In this research data mining process is used to relate this surrogate measure of traffic conditions with rear-end crash occurrence on freeways. Crash and dual loop detector data from 36.25-mile instrumented Interstate-4 corridor in Orlando (FL) are used in this study. The research problem is set up as a classification problem and separate data mining based classifiers are developed to discriminate crashes belonging to different categories from normal conditions on the freeway. Based on the models developed in this study one can identify the traffic conditions prone to rear-end crashes 5-10 minutes prior to the crash. The findings of this research are proposed to be used as a proactive traffic management system which could warn the drivers about potential rear-end crashes.

Introduction

The objective of this research is development of a framework to detect crash prone conditions in real-time. To achieve these objectives loop data collected from randomly selected non-crash locations have been used in this study along with the crash data. These data most commonly include speed, vehicle counts, and lane occupancy provided every 30 seconds by loop detectors installed beneath the freeway pavement. To establish relationships between real-time traffic data, geometric parameters, and rear-end crashes a data mining approach is adopted. It essentially means that tools from a range of fields such as machine learning (e.g., clustering algorithms), statistics (e.g., classification tree), and/or artificial intelligence are used in a step by step manner to analyze the data.

This research is part of a new trend in freeway traffic management which until recently was focused on timely detection of incidents. With the enormous increase in mobile phone usage in the recent past relevance of incident detection is diminishing and traffic management authorities are looking for proactive strategies. The basic element of a proactive traffic management system would be reliable models separating crash prone conditions from 'normal' traffic conditions in real-time. Most of the existing real-time crash 'prediction' models available in the literature are generic in nature, i.e., single generic model has been used to identify all crashes (such as rear-end, sideswipe, or angle). Conditions preceding crashes are likely to differ by type of crash and therefore the approach towards proactive traffic management should be type (of crash) specific in nature. The disaggregate models would also be useful in devising specific countermeasures for crashes. In this research the focus is on the most frequent group of crashes on the freeways, i.e., the rear-end crashes. The rear-end crash data for this study are collected over a five year period (1999 through 2003) from 36.25-mile corridor of Interstate-4 in Orlando metropolitan area along with information about geometric design features, such as ramp locations, curvature, etc. The corridor has a total of 69 loop detector stations in each direction, spaced out at nearly half a mile. Each of these stations consists of dual loops and measures average speed, occupancy, and volume over 30-second period on each of the three through travel lanes in both directions.

Background

Lee *et al.* (2002, 2003) developed and refined log-linear models to predict crashes using crash precursors estimated from loop detector data. It was found that the coefficient of temporal variation in speed has a relatively longer-term effect on crash potential than density while the effect of average variation of speed across adjacent lanes was found to be insignificant. A study by Oh *et al.* (2001) also showed the 5-minute standard deviation of speed value to be the best indicator of 'disruptive' traffic flow leading to a crash as opposed to 'normal' traffic flow. Garber and Subramanyan (2002) demonstrated the feasibility of developing a methodology in which real-time data can be used to formulate traffic management strategies also incorporating crash risk. In our previous work (Abdel-Aty et al., 2004, 2005) case-control logistic regression models were developed with matched sampling of traffic flow characteristics for crash and non-crash cases while controlling for other external factors such as the roadway geometry, time of the day, etc. These generic logistic regression models achieved satisfactory classification accuracy.

The major shortcoming of these studies was that the inferences were made based on a 'one-size-fits-all' approach. It means all types of crashes were sought to be identified using a single generic model. In this study we try to overcome these deficiencies by examining traffic data from a series of loop detectors in order to explore their relationship with rear-end crashes. The choice of rear-end crashes was obvious due to their high frequency and significant impact on freeway operation. Loop data belonging to rear-end crashes have been used with non-crash data that are collected from randomly chosen corridor locations over the 5-year period (1999 through 2003). The random sampling of non-crash locations enables us to explore the impact of 'off-line' factors (e. g, presence of ramps, time of day, horizontal curvature), along with real-time traffic parameters, on occurrence of a rear-end crash.

Data Collection

There were *2179* rear-end crashes reported in the study area during the five year period (from 1999 through 2003). From the FDOT (Florida Department of Transportation) crash database we extracted information such as the date and mile-post location for each crash. Scanned copies of individual crash reports were then used to extract the reported time of the crashes. Based on a shock-wave progression based methodology and the precise location of the crash (known from the FDOT crash database) we ascertained that the reported time was in fact very close to the time of occurrence (Pande, 2005). Loop data corresponding to crashes would be used to train the models about the crash prone conditions while a sample of randomly selected non-crash cases would be used to 'teach' the model about what constitutes 'normal' freeway traffic.

Loop data collection for crash and non-crash cases

Loop data were extracted for every crash in a specific format, for example, if a crash occurred on April 12, 1999 6:00 PM, I-4 Eastbound and the nearest loop detector was at station 30, data were extracted from station 30, two loops upstream and two loops downstream of Station 30 for 20-minute period prior to the reported time of the crash. Hence, this crash case will have loop data table consisting of the 30-seconds averages of speed, volume, and occupancy for all three lanes at stations 28 through 32 (on eastbound direction) from 5:40 PM to 6:00 PM on April 12, 1999. The choice of four stations and 20 minutes was based on results from our previous studies (Abdel-Aty et al., 2004, 2005).

The raw 30-second data have random noise and are difficult to work with in a modeling framework. Therefore, the 30-second raw data were combined into 5-minute level in order to obtain averages and standard deviations. For 5-minute aggregation 20-minute period was divided into four time slices. The stations were named as "*D*" to "*H*", with "*D*" being farthest station upstream and so on. It should be noted that "*F*" is the station closest to the location of the crash (Station of the crash) with "*G*" and "*H*" being the stations downstream of the crash location. Similarly the 5-minute intervals were also given "IDs" from 1 to 4. The interval between time of the crash and 5 minutes prior to the crash was named as time-slice 1, interval between 5 to 10 minutes prior to the crash as time-slice 2, and so on. The parameters were also aggregated across the three lanes and the averages (and standard deviations) for speed, volume, and lane-occupancy at 5-minute level were calculated based on 30 (10*3 lanes) observations. It is worth mentioning that if at a freeway location the detector from a certain lane reported missing/invalid data, the observations from that lane were not used for calculating averages and standard deviations. In such scenario (with missing/invalid data from one or two out of the three lanes) there would be less observations (either 10 from one lane or 20 from two lanes) available to get a measure of traffic flow at that location.

The format of the traffic data collected with respect to time and location of crashes and the nomenclature for independent variables is shown in Figure 1. The variable "*SSD2*", for example, represents the standard deviation of 30 speed observations during the 5-minute period of 5-10 minutes prior to a crash at station "D", which is the farthest upstream station. Note that, due to random intermittent failure of certain detectors, traffic data were not available for all *2179* rear-end crashes. Hence, the analysis presented in this paper is based on 1620 crashes which had the corresponding loop data available after disregarding the cases with some missing loop data.

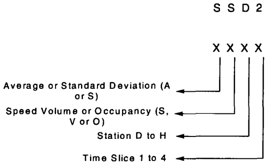
S S D 2
X X X X
Average or Standard Deviation (A or S)
Speed Volume or Occupancy (S, V or O)
Station D to H
Time Slice 1 to 4

Figure 1. Nomenclature of independent variables with respect to time and location of the crash

As mentioned earlier, the sample of non-crash cases was generated randomly from the loop database consisting of traffic data from the year 1999 through 2003. The sample was expectedly almost uniform over the times of day, day of week, and freeway location. The details of the process of generating the sample may be found in Pande (2005). These random combinations were extracted in the same format as the crash data, i.e., sets of 20-minute loop data prior to the assigned time of the non-crash from 2 stations upstream and 2 stations downstream of a station assigned as station of non-crash. The random non-crash data were aggregated to 5-minute level and traffic parameters similar to crash cases (refer to Figure 1) were generated for this sample as well. The variable "y" was given the value 0 for these cases. After the assembly of traffic parameters, geometric features of the freeway at the locations of aforementioned crash and non-crash cases were collected. This information includes the distances of nearest on and off-ramp from crash (and non-crash) locations, in both upstream and downstream direction and horizontal curvature of the freeway. The database assembled in this study is by far the most comprehensive database for development of a proactive traffic management strategy.

Data Analysis

As part of preliminary analysis, distributions of average speeds just before the crash were examined at loop detector locations surrounding the crash location. The histogram distributions for variables *ASD1, ASF1*, and *ASH1* (5-minute average speeds at Station D, F, and H, respectively) over all rear-end crashes appear to have the shape of two adjacent approximately mound-shaped distributions. These distributions suggest that the crashes belonging to each peak need to be analyzed separately. In one of our previous studies (Abdel-Aty et al., 2005), crashes were separated by simply splitting the crash data based on the average speeds at station F just before the crash (time slice 1, 0-5 minutes before the crash). In this analysis, the idea of separating crashes by prevailing conditions only at station of the crash (station F) is refined. It is imperative because rear-end crashes at freeway locations are expected to be affected not only by the prevailing speeds at that location but also by the interplay between traffic speeds at the locations upstream and/or downstream of it. To reflect this fact, it was decided to cluster the rear-end crashes into two segments/clusters/groups, based not only on *ASF1* but also on traffic speeds measured at the extremities of the 2-mile stretch around crash location (i.e., *ASD1, ASF1* and *ASH1*).

Kohonen vector quantization (*KVQ*) technique (Kohonen, 1988) was used to cluster the crash data into two groups with three average speed parameters *ASD1, ASF1,* and *ASH1* as inputs. It is intended that separate models will be applied to predict the two groups (segments/clusters) of rear-end crashes. From an application perspective, one must be able to identify the cluster to which the real-time data under consideration belong, so that appropriate model(s) may be applied to assess whether or not it is a crash prone pattern. It can not be achieved through an unsupervised learning algorithm such as the *KVQ* method. Therefore, a set of classification rules were needed that may be used to assign real-time traffic speed patterns into one of the two clusters. Classification tree was selected as the tool to formulate these rules. The rules formulated by the classification tree model to separate rear-end crashes belonging to one cluster from the other are summarized in Table 1. Note that although the clusters in rear-end crashes were obtained based on traffic speeds prevailing right before the crash (0-5 minutes; time-slice 1) the rules in Table 1 use average traffic speeds from time-slice 2 (5-10 minutes before the crash). In a real-time application, it would allow more leverage in terms of time available to analyze the data before the crash actually occurs.

TABLE 1: The series of rules to identify clusters in rear-end crash data

Leaf	Conditions (Series of Rules)	Cluster Assigned
1	ASF2<44.146 and ASD2<51.26	cluster 1
2	ASF2<44.146 and ASD2>=51.26 and ASH2 <46.8	cluster 1
4	ASF2>=44.146 and ASH2<32.941 and ASD2<53.165	cluster 1
6	ASF2>44.146 and ASH2>=32.941 and ASD2<27.30	cluster 1
3	ASF2<44.146 and ASD2>=51.26 and ASH2 >=46.8	cluster 2
5	ASF2>44.146 and ASH2<32.941 and ASD2>=53.165	cluster 2
7	ASF2>44.146 and ASH2>=32.941 and ASD2>=27.30	cluster 2

From these rules it may be inferred that the cluster 1 rear-end crashes generally belong to low speed traffic regime, while those in cluster 2 belong to medium to high speed traffic regime. Hence, most of cluster 2 crashes occur under relatively free flow conditions that commonly prevail on freeways. Based on these observations one could infer that cluster 1 rear-end crashes occur during congested conditions that prevail on the freeway for small part of the day and have very low exposure. Based on these classification rules about 45.8% crashes were identified as cluster 1 while the remaining 53.8% were identified as cluster 2. If we apply classification rules from Table 1 to a dataset with random non-crash cases then only 6.27% were classified as cluster 1. It means that although cluster 1 makeup 45.8% of the crash dataset, it only makes 6.27% of the random non-crash sample. It indicates that the crashes belonging to cluster 1 may be 'predicted' (or anticipated) using the classification tree rules shown in Table 1. If we assign all traffic patterns belonging to cluster 1 as rear-end crashes, we would be able to identify about 46% of rear-end crashes by issuing warnings just over 6% of the times. Same procedure, however, would not work for cluster 2 rear-end crashes since cluster 2 traffic conditions are way more frequent (94% in the randomly selected loop data patterns) on the freeway. Hence, further classification models are needed to separate crashes from the non-crash cases within the traffic data belonging to cluster 2.

Classification models for cluster 2 rear-end crashes

In this section the data mining process is extended to develop classification model(s) for cluster 2 rear-end crashes. Classification models developed for cluster 2 rear-end crashes belong to multi-layer perceptron (MLP) and normalized radial basis function (NRBF) neural network architecture. Theoretical details of the two architectures and training procedures may be found in any standard neural network text, e.g., Christodoulou and Georgiopoulos (2001). Inputs to the classification models are decided based on a classification tree based variable selection procedure developed by Brieman et al. (1984). Details of the procedure and the variables included may be found in Pande (2005).

The neural network modeling was repeated in three steps. In first step, the independent variables included were the off-line factors and the traffic parameters measured only at station nearest to the crash location (i. e., Station F). In the next step, traffic parameters were included from three stations, station of crash and one station each in the upstream and downstream direction. In the third step traffic parameters were included from five stations, i.e., Station D through H. The input dataset has 878 cluster 2 rear-end crashes along with the non-crash cases which made 85% of the sample. The most critical parameter affecting the performance of neural networks is the number of nodes in the hidden layer (Cybenko, 1986). To select appropriate number of nodes in the hidden layer, the performance of ten different networks with hidden nodes varying from 1 through 10 were examined for *MLP* as well as *NRBF* architecture using Enterprise Miner from SAS Institute (SAS Institute, 2001).

To evaluate the performance of neural networks these models were applied to the validation dataset. The output of these models (for any observation) is the posterior probability (0<posterior probability<1) of the event of interest (i.e., a rear-end crash). The closer it is to unity the more likely, according to the model, it is for that observation to be a rear-end crash. 30% observations with maximum posterior probability are classified as crash and various models were examined based on the proportion of the validation dataset crashes captured within those observations.

NRBF with four hidden neurons were found to be best models when traffic parameters from one and three stations were included inputs. MLP with 8 hidden neurons were the best model when traffic parameters from 5 stations were used as inputs. The performance of the models was improved by combining the best models in each category. To combine two or all three of the best models in each category the output posterior probability was averaged for the individual models. It was found that hybrid model created by combination of the three models performed better than the individual model and identified 55.4% of crashes in the validation dataset. The combination of best 1-station and 3-station model identified 53.05% crashes and the best 1-station models was the NRBF model with 4 hidden neurons, which identified 50.06% crashes from the validation dataset within the 30% observations with maximum posterior probability.

It is worth mentioning at this point that 55.4% identification of cluster 2 crashes was achieved through the model that uses traffic data from five stations (combination of best 1-station, 3-station, and 5-station models). Since data from five stations may not be simultaneously available due to intermittent failure of loops; performance of the models must be seen in terms of their data requirements as well. Sometimes it would be more practical just to use data from one station to identify these crashes. Therefore, even though the model provides better identification of cluster 2 crashes it would not make it an automatic choice for field implementation.

Real-time identification of rear-end crashes

Based on the data analysis presented here a real-time application strategy to identify conditions prone to rear-end crashes may be formulated. The strategy may be used to flag locations which are experiencing high risk of rear-end crashes. The application first starts by applying classification tree model based rules shown in Table 1. Those rules may be used to identify whether traffic data belong to cluster 1 or cluster 2. If the patterns belong to cluster 1 a rear-end crash warning is issued for the location without any further application. If the patterns are identified to be cluster 2 then we need to apply the neural network based hybrid models. As mentioned earlier, the hybrid models that combines best 1-station, 3-staton, and 5-station MLP/NRBF models provided optimal crash identification over the validation dataset and hence is preferred over other models. This model, of course, would need data from five stations around the section where we are trying to assess the crash risk. Therefore, in the next step check for data availability over five stations is applied. If data from five stations are available then the data are subjected to the hybrid model. If the requisite data are not available then the patterns may be subjected to the models with less accuracy but more tolerant data requirements (requiring inputs from 1 or 3 loop detector stations).

With this strategy one can identify *46%* of rear-end crashes (percentage of cluster 1 crashes among all rear-end crashes) by issuing warnings for about 7% cases. 55.4% of cluster 2 crashes, which make 54% of the rear-end crash data, may be identified by issuing warnings 30% of the times among the remaining 93% cases. It essentially means that about ¾ (46 + (54*55)/100 ≈ 75%) of the crashes could be identified by issuing warnings for about one third of cases (7 + (93*30)/100 ≈ 34%). It roughly translates into 66% accuracy on non-crash data for identification of 75% crashes. Further research is recommended to estimate the frequency and the nature of these warnings.

Conclusions

The paper presents a step by step approach of data analysis to develop a strategy to identify real-time traffic conditions prone to rear-end crashes using freeway loop detector data. It was concluded that the rear-end crashes on the freeway may be grouped into two distinct clusters based on the average speeds prevailing in approximately 2-mile section around the crash location 5-10 minutes before a crash. One cluster (group) of crashes occurs under extended congestion on the freeway while the average speeds are relatively higher during the 5-10 minute period before a cluster 2 crash (refer Table 1 for specific traffic speed conditions for each group of rear-end crashes). It was noticed that conditions belonging to cluster 1 occur very rarely and hence whenever such conditions are encountered in real-time then a crash warning may be issued. For cluster 2 rear-end crashes further neural network based classification models were developed. Based on the performance of the classification models and the proposed real-time application strategy, 75% of the rear-end crashes may be identified 5-10 minutes before their occurrence with just 34% positive decisions (i.e., crash warnings). Since crashes (however frequent on the I-4 corridor under consideration) are rare events; these positive decisions would result in a significant number of 'false alarms'. However, it should be noted that 'false alarms' are not as detrimental in the present application as they are for incident detection algorithms. Crash prone traffic conditions, which have been identified in this paper, would not always result in a rear-end crash occurrence. The conditions, however, are worth warning the drivers and drivers need to be more attentive under such traffic conditions even if they may not always culminate in a rear-end crash. A reasonable number of warnings, which the drivers do not consider excessive, based on the models developed can potentially play a critical role in proactive traffic management. These warnings may be issued to the motorists driving on the freeway locations through VMS (variable message signs). However, the frequency and impacts of such warnings on driver behavior would need to be carefully estimated before implementing such measure. Another application for the findings of this research could be the formulations of VSL (variable speed limit) implementation strategies that can reduce the probability of rear-end crashes.

References

Abdel-Aty, M., Uddin, N., Abdalla, F., Pande, A., and Hsia, L. (2004). Predicting freeway crashes based on loop detector data using matched case-control logistic regression. *Transportation Research Record* 1897, pp. 88-95.

Abdel-Aty, M., Uddin, N., and Pande, A. (2005). Split models for predicting multi-vehicle crashes under high speed and low speed operation conditions on freeways. *Transportation Research Record* 1908, pp. 51-58.

Breiman, L., Friedman, J., H., Olshen, R., A., and Stone, C., J. (1984). Classification and regression trees. *Chapman and Hall,* New York.

Christodoulou, C., and Georgiopoulos, M. (2001). Applications of Neural Networks in Electromagnetics, *Artech House*, Boston.

Cybenko, C. (1989). Approximations by superposition of sigmoid functions. *Mathematics of Control Signals and Systems, Vol. 2*, pp. 303-314.

Garber, N.J., Subramanyan, S. (2002). Feasibility of incorporating crash risk in developing congestion mitigation measures for interstate highways: a case study of the Hampton roads area. Report No. FHWA/VTRC 02-R17. Virginia Transportation Research Council, Charlottesville, VA.

Hand, D., Mannila, H., and Smyth, P. (2001). Principles of data mining. *The MIT Press,* Cambridge, Massachusetts.

SAS Institute, (2001). Getting Started with Enterprise Miner Software, *Release 4.1, SAS Institute*, Cary, NC.

Kohonen, T. (1988). Learning vector quantization, *Neural Networks, 1 (suppl 1),* 303.

Lee, C., Saccomanno, F., and Hellinga, B. (2002). Analysis of crash precursors on instrumented freeways. *Transportation Research Record 1784*, pp. 1-8.

Lee, C., Hellinga, B., and Saccomanno, F. (2003). Real-time crash prediction model for the application to crash prevention in freeway traffic. *Transportation Research Record 1840*, pp. 67-78.

Oh, C., Oh, J., Ritchie, S., and Chang, M. (2001). Real-time estimation of freeway accident likelihood. Presented at the *80th annual meeting of Transportation Research Board*, Washington, D.C.

Pande, A., Applying hybrid models for real-time crash risk assessment on freeways. Ph. D. Dissertation, *University of Central Florida*, 2005.