Proceedings of the 8th International
IEEE Conference on Intelligent Transportation Systems
Vienna, Austria, September 13-16, 2005

Identification of Rear-end Crash Patterns on Instrumented Freeways: A Data Mining Approach

A. Pande and M. 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 (data from freeway loop detectors) with occurrence of rear-end crashes on freeways. The results from this analysis are envisioned to be the first step in the development of a functional proactive traffic management system.**

The dataset under consideration includes information on crashes and corresponding traffic data collected from detectors neighboring the crash locations just prior to the time of the crash. The problem is setup as a classification problem for a crash being rear-end vs. not. Three types of classification tree involving different splitting criterion were attempted for variable selection. It was found that the classification tree with chi sq. test as the splitting criterion resulted in the most inclusive list of variables. The variable selection was followed by two neural network architectures, namely, the RBF (radial basis function) and MLP (multi-layer perceptron) to model the binary target variable. The two neural network models were then combined based on their output to achieve any possible improvement in the classification accuracy. It was found, however, that the classification tree model with chi sq. test as splitting criterion (with more than 65% classification accuracy) was better than any of the individual or combined neural network models (54-55% classification accuracy). Since the decision tree model also provides simple interpretable rules to classify the data in a real-time application it was recommended as the final classification model.

Manuscript received February 21, 2005. This work was supported in part by the Florida Department of Transportation.

Mohamed Abdel-Aty is with the Department of Civil and Environmental Engineering, University of Central Florida, Orlando, FL 32826 USA (Corresponding author Phone: 407-823-5657; fax: 407-823-3315; e-mail: mabdel@mail.ucf.edu).

Anurag Pande is with the Department of Civil and Environmental Engineering, University of Central Florida, Orlando, FL 32826 USA (e-mail: anurag@mail.ucf.edu).

I. Introduction

Research in the field of freeway traffic management has been mainly focused on timely detection of incidents to minimize their impact on freeway operation. However, with enormous increase in cell phone usage relevance of incident detection is diminishing and traffic management authorities are becoming more interested in pursuing proactive traffic management strategies. Of all the incidents crashes are arguably of the most critical and "predictable" type. The essential idea of a fully functional proactive traffic management system would involve anticipating incidents, such as the crashes, prior to their occurrence and then intervene in a certain manner to reduce their likelihood. The shifting of focus on to proactive traffic management has recently led to some research efforts aimed at developing crash "prediction" models. However, these models are largely generic in nature, i.e., one generic model has been used to predict different types (such as the rear-end sideswipe, or angle) of crashes. This "one size fits all" approach is of course not sufficient because different types of crashes have been known to be related to distinct traffic flow characteristics [1].

While the traffic conditions following crashes of different types (such as rear-end, sideswipe or angle crashes) are similar in nature; the conditions preceding them are likely to differ from type to type. E.g., the rear-end crashes might be expected to occur under congested traffic regime where the drivers have to slow down and speed up quite often, on the other hand the single vehicle crashes might result from excessive speeds on a curved freeway section. Therefore, while generic models may be used to separate post-incident traffic surveillance data from a non-incident scenario; the approach for proactive traffic management should be type (of crash) specific in nature. Even though the eventual goal might be to estimate models that would separate conditions prone to a certain type of crash from non-crash conditions; a set of rules/models should first be devised to decide about models belonging to which specific type(s) of crashes should come into play under the existing traffic conditions. Hence, the identification of the most probable type of crash under a traffic scenario would be the first step required for development of a proactive system. Such models/rules would also be useful while devising remedial measures to improve the safety situation on the freeway which would differ for each type of crash, e. g, the variable speed limits for rear-end crashes or a temporary "no lane-changing" sign to avoid an impending sideswipe crash.

0-7803-9215-9/05/$20.00 ©2005 IEEE.

In this paper a data mining approach is proposed to separate rear-end crashes from other types based on freeway traffic data collected through the loop detector stations surrounding the location of historical crashes. The choice of rear-end crashes was obvious since these are the crashes most frequent on the freeway facilities and make up a little more than 50% of our crash data.

Data for this study were collected from 36.25-mile instrumented corridor of Interstate-4 in the central Florida area. The information about historical crashes that occurred on the freeway during the five-year period was collected from the FDOT (Florida Department of Transportation) intranet server and the corresponding traffic related variables were extracted from the loop detector database previously maintained at University of Central Florida. The formation and structure of the dataset would be discussed in detail later in the paper. The data mining process involving data preparation, data partition, variable selection, model building, and assessment, was implemented using Enterprise Miner from SAS Institute [2].

II. BACKGROUND

Madanat and Liu [3] came up with an incident likelihood prediction model using loop data as input. The focus of their research was to enhance existing incident detection algorithms with likelihood of incidents. They actually considered two types of incidents a) crashes and b) overheating vehicles. They concluded that merging section, visibility and rain are statistically the most significant factors for crash likelihood prediction.

Lee et al. [4, 5] developed and refined log-linear models to predict crashes through estimation of crash precursors 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.

Oh et al. [6] and Abdel-Aty and Pande [7] developed density estimation based models to classify the pre-crash temporal variation in speed into crash vs. non-crash.

The authors in their earlier studies [8, 9] developed logistic regression model that utilized information on traffic flow characteristics for crash and matched non-crash cases while controlling for other external factors (thereby implicitly accounting for factors such as the geometry and location). In one of the more detailed study, Golob and Recker [1] concluded that the collision type is the best-explained characteristic and is related to the median speed and left-lane and interior lane variations in speed. They also pointed out that some collision types are more common under certain existing traffic conditions.

It must be noted that the models developed in all these studies, with the exception of [1], were generic in nature, i.e., one model was developed to separate crashes from non-crash cases irrespective of their collision type. However, the findings from these studies are still useful for us since the rear-end crashes make up majority of freeway crashes and any generic mode would tend to be biased toward identifying the traffic factors associated with rear-end crashes. In that sense the contribution of these studies towards proactive traffic management is obviously significant.

In this paper a data mining approach is presented to analyze the crash and corresponding loop detector data to differentiate rear-end crashes from those of the other types (i.e., sideswipe, angle crashes etc.). Due to emergence of very large databases and computer automated data recording in science and engineering, the level of interest in data mining has increased significantly. Data mining sits at the common frontiers of several fields including database management, artificial intelligence, machine learning, pattern recognition, and data visualization [10]. Although certain data mining tools such as the classification tree, MLP and RBF neural networks have been individually employed in the area of incident detection and traffic safety [e.g., 11, 12] their application in a data mining process framework has been almost non-existent in traffic management research.

Data mining procedures are usually applied in an "observational" setting rather than an "experimental" setting. It means data mining typically deals with data that have been already been collected for some purpose other than the data mining analysis. This is one way in which data mining sharply differs from traditional statistics, where data are often collected by using efficient strategies to answer specific questions (experimental design) [13].

The idea of using the loop detector data for proactive traffic management and traffic safety research by linking it to crash patterns falls in the former category of observational setting. With huge amounts of ITS-related data being archived for applications such as the travel time prediction etc., data mining process is suitable for relating this huge amount of data to specific crash patterns.

III. DATA PREPARATION

Traffic surveillance data collected through underground dual loop detectors on Interstate-4 (I-4) are used in this study. These detectors record and archive following traffic flow parameters every 30 seconds: average vehicle counts, average speed, and lane detector occupancy (percentage of time the loop is occupied by vehicles). These data are collected from three lanes in each direction through 69 stations spaced at approximately 0.8 km (0.5 mile) for a 58-km (36-mile) stretch in each direction. A typical dual loop detector system along with its spatial arrangement on the Eastbound I-4 segment is shown in Figure 1.

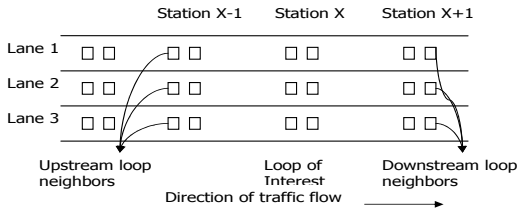

Figure 1: Configuration of loop detectors on the freeway segment

The crash data for the study were collected from the FDOT crash database for the years 1999 through 2003. Besides date, time and location of each crash the database provided details on characteristics such as type and severity of crashes. Based on information from one of the variables "first_harmful_event" available in the FDOT crash database binary variable named "rear" was created. The variable "first_harmful_event" correspond to the type of crash. The variable "rear" was defined as 1 if the "first_harmful_event" was a rear-end collision and 0, otherwise.

The location for each crash that occurred in the study area during the period of analysis was then identified. For every crash, the loop detector station nearest to its location was determined and referred to as the station of the crash. The pre-crash loop detector data from stations surrounding the crash location were collected based on the reported time of historical crashes. Traffic data corresponding to the day of crash were extracted in a specific format. The correspondence here means that, for example, if a crash occurred on April 12, 2001 (Monday) 6:00 PM, I-4 Eastbound and the nearest loop detector was at station 30, data were extracted from station 30, three loops upstream and three loops downstream of station 30 for half an hour period prior to the estimated time of the crash. Hence, this crash will have the raw loop data table consisting of the speed, volume and occupancy values for all three lanes from the loop stations 27-33 (on eastbound direction) from 5:30 PM to 6:00 PM for the day of crash.

Out of little more than 4000 crashes in the sample, 52% of them were rear-end crashes. Therefore, the dataset is somewhat `balanced in terms of the target variable "rear".

A. Data Aggregation

The raw 30-second data obtained directly from loop detector have random noise and are difficult to work with in a modeling framework. Moreover, the raw loop data also suffers from auto-correlation. Therefore, the 30-second raw data was combined into 5-minute level in order to get averages and standard deviations. Thus for 5-minute level aggregation half an hour period was divided into 6 time slices. The stations were named as "C" to "I", with "C" being farthest station upstream and so on. It may be noted that "F" is the station of the crash with "G", "H" and "I" being the stations downstream of the crash location. Similarly the 5-minute intervals were given "IDs" from 1 to 6. The interval between time of the crash and 5 minutes prior to the crash was named as slice 1, interval between 5 to 10 minutes prior to the crash as slice 2, interval between 10 to 15 minutes prior to the crash as slice 3 and so on.

The parameters were further 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. Therefore, even if at a location the loop detector from a certain lane was not reporting data, there were observations available to get a measure of traffic flow at that location. Aggregating data across the lanes helps to develop a system for more realistic application scenario since all three lanes at a loop detector stations are less likely to be simultaneously unavailable when the model is used for real-time prediction. Another advantage is that the measures aggregated across lanes not only capture temporal variations (or lack there of) but variations across the three lanes as well. The format of the traffic data collected with respect to time and location of crashes is provided in Figure 2. The figure also shows the description of field nomenclature.

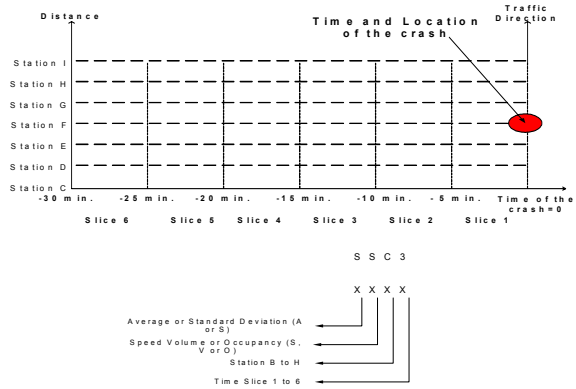

Figure 2: Traffic data collection in a time-space framework and nomenclature of independent variables with respect to time and location of the crash

The variable shown for example SSC3 represents the standard deviation in speed during the 5-minute period of 10-15 minutes prior to a crash at station "C" which is the farthest upstream station.

From our previous work [8, 9] it is known that the 5-minute coefficient of variation in speed (standard deviation of speed/average speed) observed at stations neighboring the crash location is associated with the risk of crash occurrence. The analysis in that paper was inclusive of all different types of crashes but since the rear-end crashes are more than 50% of all crashes the variables found significant in that analysis were expected to be important identifiers of rear-end crashes. Therefore the averages and standard deviation of speeds were replaced with the coefficient of variation in speed using the transformation node in the Enterprise Miner. The variables were named as "CVSXY" with the last two letters signifying the station and the time slice with which the parameters were associated, respectively.

IV. MODELING METHODOLOGY, PROCEDURE AND RESULTS

A. Modeling Methodology

SAS Institute [2] defines data mining as the process of **S**electing, **E**xploring, **M**odifying, **M**odeling, and **A**ssessing (SEMMA) large amounts of data to uncover previously unknown patterns that can be utilized for business advantage. In this paper these steps are followed to develop classification models separating the loop detector data patterns preceding a rear-end crash from those preceding crashes of other types. Enterprise Miner software from SAS Institute is used to implement aforementioned SEMMA data

mining process. The SEMMA process may be controlled through a flow diagram which may be modified or saved using Enterprise Miner GUI [2].

B. Modeling Issues

SAS Enterprise Miner contains a collection of sophisticated analysis and data preparation tool nodes with a common user-friendly interface. Data preparation tools include outlier detection, variable transformations, data imputation, random sampling, and partitioning of data sets (into train, test, and validate data sets). Miner may be conveniently used to create, compare and ensemble multiple models. Modeling tools include decision trees, regression, and neural networking. The performances of various models may be assessed through the Assessment Node using plots such as the ROC curve, lift chart etc. [2].

With so many options available the selection of the tool(s) to be used was a critical issue. The research problem is formulated as a classification problem and the outcome of interest is a crash being of the rear-end type (with target variable rear=1).

The loop detectors are spaced about ½-mile (0.8 km) on the freeway and provide a 30-second snapshot of the current traffic scenario. This type of data would not provide us with enough resolution to identify the causal factors or exact mechanism responsible for individual crashes. To understand the mechanism of crashes one needs detailed vehicle to vehicle movement data, which being impossible to obtain; the loop detector data is being used as a surrogate. Essentially we are trying to identify if the patterns in the data collected from loop detectors at fixed locations are leading to crash occurrences of a specific type or not. The application therefore directs us away from random sampling logistic regression models.

The classification trees are unstable modeling tool and are usually recommended for variable selection. Brieman et al. [14] devised a variable importance measure for trees. Variable importance measure may be used as a criterion to select a promising subset of variables for other flexible modeling tools such as the neural networks. The theoretical details of this measure may be found in the relevant reference [14]. As a data preparation tool the tree also offer interpretability, no strict assumptions concerning the functional form of the model and computational efficiency. At this point the neural networks were chosen as the tool for final classification due to their flexibility. Two different types of neural network architectures were examined; the multi-layer perceptron (MLP) and the radial basis function (RBF) neural network. The theoretical details of these tools may be found in any standard neural network text such as [15].

It was also decided to examine parameters from one time slice at a time in one model. It will not only avoid the autocorrelation problems but would also lead to an easy practical implementation plan. Using data from the same time duration would be easier than to collect data and wait for the model estimation until after the data from next time slice is recorded. Hence the models in this paper are based on the parameters calculated between 5-10 minutes before the time of crash (i.e., parameters from time slice 2). Time slice 1 being too close to the time of crash, would allow no leverage in terms of time to use the results of the models in a proactive traffic management system. Therefore, next closest time slice (time slice 2) is used here. This leaves us with 35 candidate variables (averages and standard deviation of volume and occupancy and all 7 stations around the crash location from which data is extracted 7*2*2=28 and 7 coefficients of variation in speed; 28+7=35) belonging to time slice 2.

C. Modeling Procedure and Results

As the first step in modeling process the dataset was split into training and validation samples through the data partition node using simple random sampling. Standard 2:1 split was used for training and validation, respectively. The final data mining process flow diagram from SAS Enterprise Miner is shown in Figure 3.

It may be seen that the data partition node is followed by three separate tree nodes attempted for variable selection. The three tree nodes use different splitting criterion, namely, the chi-sq. test, entropy reduction and gini measure of impurity reduction. The best split among available set of candidate splits is determined using these criterions. The tree nodes are used here to identify the important variables from the aforementioned 35 candidate variables belonging to time slice.

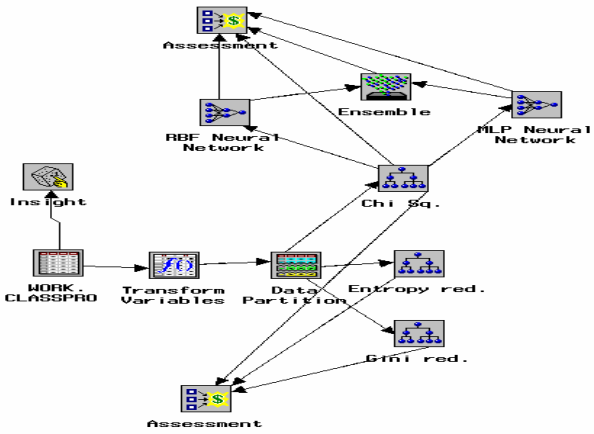

Figure 3: Data mining process flow diagram

The variable importance measures (devised by Brieman et al. [14]) based on each of the three possible splitting criterion were calculated for every variable using the three tree nodes and only the variables having importance measure greater than 0.05 were to be retained for further use in the two neural network architectures. Out of the three different list of variables generated by each of the tree node it was decided to select the output of the tree resulting in the most interpretable set of important variables to use in the next step (model building) of modeling procedure. Note that the purpose of the tree is variable selection then the number of surrogate splits should be increased from the Enterprise Miner default value zero. Keeping the default value

unchanged might result in exclusion of some important variables.

The variables identified by the three tree nodes are shown in Table 1. The leftmost column shows the variables selected by the tree using chi sq. criterion followed on the right by the lists of variables selected by the trees using entropy reduction and gini reduction criterions, respectively.

List of variables selected through tree model using		
Chi Sq. split criterion	Entropy reduction split criterion	Gini measure of impurity reduction split criterion
CVSH2, AOH2	CVSH2	CVSH2
CVSG2, AOG2	CVSG2	CVSG2
CVSF2, AOF2	CVSD2	SOG2
CVSE2, AOE2	AOG2	AOG2
CVSD2, AOD2	AOF2	AOF2
SOH2, AOC2	AOE2	AOD2
SOG2, SVG2	SVH2	AOC2
SOF2, AVH2	AVG2	SVG2
SOE2, AVG2, AVF2, AVD2, AVC2	AVC2	AVG2, AVC2

Table 1: List of variables selected by the separate tree models using different splitting criterion, namely, chi sq. test, entropy reduction and gini reduction

It may be seen that chi sq. tree resulted in a relatively exhaustive list of variables selected. From an interpretation point of view the list of variables selected by this was more inclusive. Since this is a preliminary step in modeling and we did not want to risk loosing any critical variable it was decided to go along with the tree using the chi sq. splitting criterion. It may be observed in the list that the CVS (coefficient of variation in speed) and SO (Standard deviation of occupancy) at upstream as well as downstream stations are one of the critical parameters associated with rear-end crashes. AO (average occupancy) and AV (average volume) are also significant. SV (standard deviation in volume) is one parameters which is only significant at downstream of crash location (Station G).

Examining the hierarchical structure of the classification tree used to obtain variable importance measure for each variable it was noticed that if AOF2 (Average occupancy at station of crash) >= 11.449; 77.1% of crashes in the validation sample were rear-end. Moreover, if AOF2>= 11.449 and CVSG2 >=1.118; about 80% of crashes in the validation sample were rear-end. These two splits point toward frequent formation and dissipation of ephemeral traffic queues under congested traffic regime characterized by high occupancy and high coefficient of variation in speed. Under such conditions the drivers have be very attentive in following other vehicles and even a little lapse in concentration could cause a rear-end crash.

It may be seen in Figure 4 that the tree node was followed by two parallel neural network nodes. The two architectures used here are the RBF (radial basis function) and MLP (multilayer perceptron) neural networks. It has been proven in the literature that an MLP structure with one hidden layer and nonlinear activation functions for the hidden nodes can implement any function of practical interest [16]. Hence, it was sensible to focus on MLP structure with one hidden layer and not complicate the structure unnecessarily. The number of neurons in the hidden layer was, however, varied from 1 through 20 and the classification performance of each model on the validation dataset was observed. It was found that the network with 12 hidden layers provided the best classification performance. The model achieved 54% classification accuracy on the validation dataset.

Similarly two types of RBF architecture with equal and unequal width were examined and it was observed that the network with unequal width provided the best classification performance (55% compared to 34% of the equal width RBF network) over the validation dataset. The misclassification rate on the validation dataset for the optimal RBF and MLP networks was 45% and 46%, respectively. The next step was to ensemble the two neural networks and check if the classification accuracy improves. It was found that combining the two models through the ensemble node, based on the average of the posterior probability obtained from the two individual models, did not improve the classification accuracy over the validation dataset. Indeed when the outputs of the two models were compared side by side, it was found that the two models mostly agreed with each other on their respective classification for most of the observations on the validation dataset.

Classification accuracy of the neural network models was in fact worse than the diagnostic tree model used earlier for variable selection. Performance of the four models (two individual neural network models, ensemble model and the diagnostic tree model) was also compared based on percentage cumulative response lift chart generated by the Assessment Node of the Enterprise Miner.

In the lift chart, the crashes in the validation dataset are sorted from left to right by posterior probability of being a rear-end crash (model output). The sorted group is lumped into ten deciles[1] along the horizontal axis. The left-most decile would be the 10% crashes most likely to be rear-end. The vertical axis represents the actual cumulative response rate within each decile. The lift chart displays the cumulative percentage response values for a baseline model and for the four predictive models. Note that the baseline model represents the proportion (52%) of target event (rear=1) in the validation sample. The performance of each model may be measured by determining how many rear-end crashes does the models capture across various deciles. For example, according to the figure 82% crashes are rear-end within top 10% observations of the tree model. The same percentage varies between 76 to 78% for the other three models depicted in Figure 4.

Hence, according to Figure 4 the diagnostic tree model developed for variable selection captures more rear-end crashes and therefore has lift plot higher than any other model. With just 35% misclassification rate on the validation sample the tree model is recommended for final

[1] Decile is defined as any of nine points that divide a distribution of ranked scores into equal intervals where each interval contains one-tenth of the scores

classification and not only for variable selection. Note that it is possible to combine the three tree models initially attempted for variable selection to improve on the classification accuracy but then the simple interpretability of the rules would be lost.

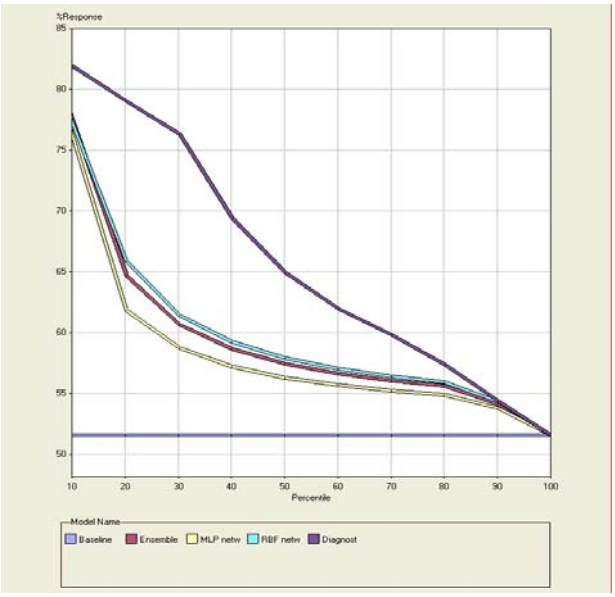

Figure 4: Lift chart showing performance of four models in terms of percentage response at various deciles

V. CONCLUSION

The paper presents a data mining approach to analyze the loop detector data preceding freeway crashes in order to identify the type of the crash (as characterized by the first harmful event associated with the crash) most likely to occur under existing traffic conditions. The focus on this paper is on the rear-end crashes that are the most frequent type on the freeways. Data mining tools classification tree, and two neural network architectures were explored in order to identify the critical factors associated with the occurrence of rear-end crashes.

To separate rear-end crashes from all other types, dataset consisting all crashes and corresponding loop data on the 36.25-mile Interstate segment was used with binary target variable "rear". It was set up as a binary classification problem in which traffic variables measured during 5-10 minutes before the crash are used as independent variable to identify crashes of the rear-end type. It was found that the tree model developed to identify the important variables was the one ultimately used for classification. Two neural network architectures (MLP and RBF) explored here did not improve on the performance of the diagnostic tree model and on the contrary did worse.

From a future application perspective, worse performance by the neural network models might be a blessing in disguise. As explained in the introduction section; identification of type of crash most likely to occur under existing traffic situation will be the first step in the process of separating non-crash data from crash prone traffic conditions. The tree model can be applied to the real-time data with simple interpretable rules and would be an ideal first component of the envisioned proactive traffic management system. Binary classification tree model(s) similar to the one developed here (for rear-end crashes) may be attempted for other common types of crashes such as side-swipe, single vehicle and angle crashes. However, the issues regarding the imbalanced sample would need to be resolved since other types of crashes are not as frequent on freeways and make up only 20% to 35% of all crash data.

Based on the simple rules the decision can be easily made by the system about prediction models of which category (e.g., etc.) to trigger. Of course we would need models capable of separating crash (rear-end, sideswipe etc.) data from non-crash data and not the ones identifying type of crash given a crash has occurred. A similar data mining based approach may be used for developing those models as well.

REFERENCES

[1] Golob, T. F., and Recker, W. W., Relationships among urban freeway accidents, traffic flow, weather and lighting Conditions. California PATH Working Paper UCB-ITS-PWP-2001-19, Institute of Transportation Studies. University of California, Berkeley, 2001.

[2] SAS Institute, Getting Started with Enterprise Miner Software, Release 4.1, SAS Institute, Cary, NC, 2000.

[3] Madanat, S., and Liu, P., A prototype system for real-time incident likelihood prediction. IDEA Project Final Report (ITS-2), Transportation Research Board, National Research Council, Washington, D.C., 1995.

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

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

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

[7] Abdel-Aty, M., and Pande, A., Classification of real-time traffic speed patterns to predict crashes on the freeways. Presented at the 83rd Annual Meeting of the Transportation Research Board (TRB), Washington D.C., 2004.

[8] Abdel-Aty, M., Uddin, N., Abdalla, F., Pande, A., and Hsia, L., Predicting freeway crashes based on loop detector data using matched case-control logistic regression. Forthcoming in the Transportation Research Record, 2004.

[9] Abdel-Aty, M., Uddin, N., and Pande, A., Split models for predicting multi-vehicle crashes under high speed and low speed operation conditions on freeways. Presented at the 84th Annual Meeting of the Transportation Research Board (TRB), Washington D.C., 2005.

[10] Friedman, J. H. 1997. Data Mining and Statistics. What's the Connection? Proc. of the 29th Symposium on the Interface: Computing Science and Statistics, Houston, Texas, 1997.

[11] Abdelwahab, H., and Abdel-Aty, M., Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record, No. 1746, 2001, pp. 6-13.

[12] Sohn, S., and Shin, H., Pattern recognition for road traffic accident severity in Korea. Ergonomics, Vol. 44, No. 1, 2001, pp. 107-117.

[13] Hand, D., Mannila, H., and Smyth, P., Principles of data mining. The MIT Press, Cambridge, Massachusetts, 2001.

[14] Breiman, L., Friedman, J., H., Olshen, R., A., and Stone, C., J. Classification and regression trees. Chapman and Hall, 1984.

[15] Haykin, S., Neural networks: A comprehensive foundation. Macmillan Publishing Company, New York, 1999.

[16] Cybenko, C., Approximations by superposition of sigmoid functions. Mathematics of Control Signals and Systems, Vol. 2, 1989, pp. 303-314.