Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia

Tibebe Beshah[1], Shawndra Hill[2]

Department of Information Science[1], Operations and Information Management Department[2]
Addis Ababa University, Ethiopia[1] , University of Pennsylvania, Philadelphia, PA[2]
tibebe.beshah@gmail.com[1], shawndra@wharton.upenn.edu[2]

Abstract

Road traffic accidents (RTAs) are a major public health concern, resulting in an estimated 1.2 million deaths and 50 million injuries worldwide each year. In the developing world, RTAs are among the leading cause of death and injury; Ethiopia in particular experiences the highest rate of such accidents. Thus, methods to reduce accident severity are of great interest to traffic agencies and the public at large. In this work, we applied data mining technologies to link recorded road characteristics to accident severity in Ethiopia, and developed a set of rules that could be used by the Ethiopian Traffic Agency to improve safety.

Problem Statement

The costs of fatalities and injuries due to road traffic accidents (RTAs) have a tremendous impact on societal well-being and socioeconomic development. RTAs are among the leading causes of death and injury worldwide, causing an estimated 1.2 million deaths and 50 million injuries each year (World Health Organization, 2004). Ethiopia has the highest rate of RTAs, owing to the fact that road transport is the major transportation system in the country. The Ethiopian traffic control system archives data on various aspects of the traffic system, such as traffic volume, concentration, and vehicle accidents. With more vehicles and traffic, the capital city of Addis Ababa takes the lion's share of the risk, with an average of 20 accidents being recorded every day and even more going unreported.

The basic hypothesis of this research is that accidents are not randomly scattered along the road network, and that drivers are not involved in accidents at random. There are complex circumstantial relationships between several characteristics (driver, road, car, etc.) and the accident occurrence. As such, one cannot improve safety without successfully relating accident frequency and severity to the causative variables (Kononov and Janson, 2002). We will attempt to extend the authors' previous work in this area by generating additional attributes and focusing on the contribution of road-related factors to accident severity in Ethiopia. This will help to identify the parts of a road that are risky, thus supporting traffic accident data analysis in decision-making processes.

The general objective of the research is to investigate the role of road-related factors in accident severity, using RTA data from Ethiopia and predictive models. Our three specific objectives include: 1) exploring the underlying variables (especially road-related ones) that impact car accident severity, 2) predicting accident severity using different data mining techniques, and 3) comparing standard classification models for this task.

Literature Review

Various studies have addressed the different aspects of RTAs, with most focusing on predicting or establishing the critical factors influencing injury severity (Chong, A. et al. 2005). Numerous data mining-related studies have been undertaken to analyze RTA data locally and globally, with results frequently varying depending on the socio-economic conditions and infrastructure of a given location.

Ossenbruggen, Pendharkar et al. (2001) used a logistic regression model to identify the prediction factors of crashes and crash-related injuries, using models to perform a risk assessment of a given region. These models included attributes describing a site by its land use activity, roadside design, use of traffic control devices, and traffic exposure. Their study illustrated that village sites were less hazardous than residential or shopping sites. Abdalla et al. (1997) also studied the relationship between casualty frequency and the distance of an accident from residential zones. Not

surprisingly, casualty frequencies were higher in accidents that occurred nearer to residential zones, possibly due to higher exposure. The casualty rates among residents from relatively deprived areas were significantly higher than those from relatively affluent areas.

Mussone et al. (1999) used neural networks to analyze vehicle accidents that occurred at intersections in Milan, Italy. These authors used feed-forward multilayer perception (MLP) with BP learning. The model had 10 input nodes for eight variables: day/night, traffic flows in the intersection, number of virtual and real conflict points, intersection type, accident type, road surface condition, and weather condition. The output node ('accident index') was calculated as the ratio between the number of accidents at a given intersection and at the most dangerous intersection. Results showed that the highest accident index for the running over of pedestrians occurred at non-signalized intersections at nighttime.

Sohn and Hyungwon (2001) conducted research on pattern recognition in the framework of RTA severity in Korea. They observed that an accurately estimated classification model for several RTA severity types as a function of related factors provided crucial information for accident prevention. Their research used three data mining techniques, neural network, logistic regression, and decision tree, to select a set of influential factors and to construct classification models for accident severity. Their three approaches were then compared in terms of classification accuracy. They found that accuracy did not differ significantly for each model, and that the protective device was the most important factor in the accident severity variation.

To analyze the relationship between RTA severity and driving environment factors, Sohn and Lee (2002) used various algorithms to improve the accuracy of individual classifiers for two RTA severity categories. Using neural network and decision tree individual classifiers, three different approaches were applied: classifier fusion based on the Dempster–Shafer algorithm, the Bayesian procedure, and logistic model; data ensemble fusion based on arcing and bagging; and clustering based on the *k*-means algorithm. Their empirical results indicated that a clustering-based classification algorithm works best for road traffic accident classification in Korea.

Ng, Hung and Wong (2002) used a combination of cluster analysis, regression analysis, and geographical information system (GIS) techniques to group homogeneous accident data, estimate the number of traffic accidents, and assess RTA risk in Hong Kong. Their resulting algorithm displayed improved accident risk estimation compared to estimates based on historical accident records alone. The algorithm was more efficient, especially for fatality and pedestrian-related accident analyses. The authors claimed that the proposed algorithm could be used to help authorities effectively identify areas with high accident risk, and serve as a reference for town planners considering road safety.

Chang and Chen (2005) conducted data mining research focusing on building tree-based models to analyze freeway accident frequency. Using the 2001-2002 accident data of National Freeway 1 in Taiwan, the authors developed classification and regression tree (CART) and negative binomial regression models to establish the empirical relationship between traffic accidents and highway geometric variables, traffic characteristics, and environmental factors. CART is a powerful tool that does not require any pre-defined underlying relationship between targets (dependent variables) and predictors (independent variables). These authors found that the average daily traffic volume and precipitation variables were the key determinants of freeway accident frequency. Furthermore, a comparison of their two models demonstrated that CART is a good alternative method for analyzing freeway accident frequencies.

Tibebe (2005) analyzed historical RTA data, including 4,658 accident records at the Addis Ababa Traffic Office, to investigate the application of data mining technology to the analysis of accident severity in Addis Ababa, Ethiopia. Using the decision tree technique and applying the KnowledgeSEEKER algorithm of the KnowledgeSTUDIO data mining tool, the developed model classified accident severity into four classes: fatal injury, serious injury, slight injury, and property-damage. Accident cause, accident type, road condition, vehicle type, light condition, road surface type, and driver age were the basic determinant variables for injury severity level. The classification accuracy of this decision tree classifier was reported to be 87.47%.

Chang and Wang (2006) applied non-parametric classification tree techniques to analyze accident data from the year 2001 for Taipei, Taiwan. A CART model was developed to establish the relationship between injury severity and driver/vehicle characteristics, highway/environment variables, and accident variables. The most important variable associated with crash severity was the vehicle type, with pedestrians, motorcyclers, and bicyclists having the highest injury risks of all driver types in the RTAs.

Using one clustering (SimpleKMeans) and three classification (J48, naïve Bayes, and One R) algorithms, Srisuriyachai (Srisuriyachai 2007) analyzed road traffic accidents in the Nakhon Pathom province of Bangkok. Considering the descriptive nature of the results and classification performance, the J48 algorithm was sufficiently useful and reliable. The outcome of the research was traffic accident profiles, which the author presented as a useful tool for evaluating RTAs in Nakhon Pathom.

Wong and Chung (2008) used a comparison of methodology approaches to identify causal factors of accident severity. Accident data were first analyzed with rough set theories to determine whether they included complete information about the circumstances of their occurrence according to an accident database. Derived circumstances were then compared. For those remaining accidents without sufficient information, logistic regression models were employed to investigate possible associations. Adopting the 2005 Taiwan single-auto-vehicle accident data set, the results indicated that accident fatality resulted from a combination of unfavorable factors, rather than from a single factor. Moreover, accidents related to rules with high or low support showed distinct features.

Following Tibebe's (2005) work, Zelalem (2009) conducted a data mining study to classify driver responsibility levels in traffic accidents in Addis Ababa. The study focused on identifying the important factors influencing the level of driver responsibility, and used the RTA dataset of the Addis Ababa Traffic Control and Investigation Department (AATCID). The WEKA data mining tool was used to build the decision tree (using the ID3 and J48 algorithms) and MLP (back propagation algorithm) predictive models. Rules representing patterns in the accident dataset were extracted from the decision tree, revealing important relationships between variables influencing a driver's level of responsibility (e.g., age, license grade, education, driving experience, and other environmental factors). The accuracies of these models were 88.24% and 91.84%, respectively, with the decision tree model found to be more appropriate for the problem type under consideration.

Getnet (2009) investigated the potential application of data mining tools to develop models supporting the identification and prediction of major driver and vehicle risk factors that cause RTAs. The research used the WEKA version 3-5-8 tool to build the decision tree (using the J48 algorithm) and rule induction (using PART algorithm) techniques. Performance of the J48 algorithm was slightly better than that of the PART algorithm. The license grade, vehicle service year, vehicle type, and experience were identified as the most important variables for predicting accident severity.

Finally, Liu (2009) developed a decision support tool for liability authentications of two-vehicle crashes, based on self-organizing feature maps (SOM) and data mining models. Although the study used a small data sample, the decision support system provided reasonably good liability attributions and references on the given cases.

Testbed

The accident record has more than forty columns (or attributes) of text, numbers, dates, and times. Among these attributes, the car plate number and driver's name were withheld by the AATCID for privacy purposes. Table 1 displays the relevant attributes (selected through feature selection) and their data types.

Table 1: Description of relevant categorical attributes

Attribute Name	Description
Subcity	Name of subcity where accident occurred.
ParticularArea	Whether the accident occurred in school or market areas.
RoadSeparation	How road segments are separated
RoadOrientation	How the road is oriented
RoadJunction	Type of road junction
RoadSurfaceType	Whether the road surface is asphalt or ground.
RoadSurfCondition	Whether the road surface is dry, muddy, or wet.
WeatherCondition	The weather condition
LightCondition	The light condition
AccidentSeverity	The severity of the accident

Experimentation

To predict accident severity, various classification models were built using decision tree, naive Bayes, and K-nearest neighbor classifiers. Decision trees are easy to build and understand, can manage both continuous and categorical variables, and can perform classification as well as regression. They automatically handle interactions between variables and identify important variables.

After assessing the data and selecting the predictive models to be used, a series of experiments were performed. Extensive data pre-processing resulted in a clean dataset containing 18,288 accidents with no missing values. The class label ('Accident Severity') had four nominal values: 'Fatal,' 'Severeinjury,' 'Slightinjury,' or 'Propertyloss.' During data exploration, different numbers of attributes were selected by different feature selection techniques.

Since WEKA's explorer generally chooses reasonable defaults, the J48 decision tree algorithm was performed using its default parameters: a confidence interval of 0.25, pruning allowed, and a minimum number of objects for a leaf of 3. Training and testing were done using ten-fold cross-validation.

In the first experiment, the 18,288-accident dataset with 10 attributes, including 9 independent variables and one dependent variable (the class-label attribute 'AccidentSeverity'), were fed to WEKA's explorer. The J48 classifier was used and an accuracy of 80.221 was achieved. In the second and third experiments, the same input, instances, and attributes were fed to WEKA. Using the naive Bayes classifier, an accuracy of 79.9967 was achieved. Using the K-nearest neighbors classifier (IBK), an accuracy of 80.8281 was achieved.

Results

All three classifiers performed similarly well with respect to the number of correctly classified instances (Table 2).

Table 2: Summary of experiments conducted

S.n	Classification Models (classifiers)	Number of correctly classified instances	Accuracy in percentage
1	Decision Tree (J48)	14,666	80.221%
2	Naive Bayes	14625	79.9967%
3	K-Nearest Neighbors	14777	80.8182%

The priors on the Property Loss class was approximately 75%, Slight Injury occurred approximately 10% of the time, Severe Injury occurred 8% of the time, and very few accidents were Fatal. Compared our prior (on Property Loss), we perform better than without having a model with respect to accuracy. However, accuracy alone does not completely describe the prediction efficiency, and other means of evaluating our predictive models are necessary. The receiver operating characteristics (ROC) curve, also known as the relative operating characteristic curve, is a comparison of two operating characteristics as the criterion changes. It can be represented by plotting the fraction of true positives (TPR = true positive rate) versus the fraction of false positives (FPR = false positive rate). An ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independent from (and prior to specifying) the cost context or class distribution. The ROC analysis is directly and naturally related to the cost/benefit analysis of diagnostic decision making. The area under the ROC curve (AUC) quantifies the overall discriminative ability of a test. An entirely random test (i.e., no better at identifying true positives than flipping a coin) has an AUC of 0.5, while a perfect test (i.e., one with zero false positives or negatives) has an AUC of 1.00.

Since the accuracies of the above models were almost identical, we used ROC curves to further evaluate our models, using 20% (3,657) of the instance data. Some of the visualizations of the threshold curves are presented below, followed by a summary of the AUCs for each class value of the target class for each model.

Table 3: Summary of the AUCs

S.n	Classification model (classifiers)	Class values	AUCs
1	Decision Tree (J48)	PropertyLoss	0.699
		SlightInjury	0.608
		Fatal	0.815
		SevereInjury	0.736
2	Naive Bayes	PropertyLoss	0.752
		SlightInjury	0.680
		Fatal	0.855
		SevereInjury	0.761
3	K-Nearest Neighbors	PropertyLoss	0.884
		SlightInjury	0.875
		Fatal	0.965
		SevereInjury	0.918

In all cases, the AUCs were significantly > 0.5, with the K-nearest neighbors model displaying AUCs closest to 1. These results indicate that all models predicted new instances well.

Knowledge Representation

A predictive model is useless if it cannot represent knowledge in a way that end users can understand. Many learning techniques look for structural descriptions ("rules") of what is learned, which can become fairly complex. These descriptions are easily understood by the end user, and explain the bases for new predictions. Classification rules are a popular alternative to decision trees in representing the structures that learning methods produce, partly because each rule seems to represent an independent "nugget" of knowledge (Witten and Frank 2000). The antecedent (or precondition) of a rule is a series of tests, similar to those at decision tree nodes. The consequent (or conclusion) defines the class(es) that apply to instances covered by that rule, or perhaps provides a probability distribution over the class(es).

PART is a class for generating a decision list in WEKA. The PART algorithm is used to represent the knowledge/pattern identified. To identify significant rules, PART was run on the accident dataset with different numbers of attributes. Ten-fold cross-validation was used for testing and the minimum number of objects in a leaf was set to twenty. Domain experts were consulted in evaluating the significance of the rules. Rules were generated based on the following attributes: 'RoadOrientation,' 'ParticularArea,' 'RoadSeparation,' 'Subcity,' 'Roadjunction,' and 'AccidentSeverity' (dependent) (Fig. 1). The accuracy of the algorithm in generating the rules was 79.942.

PART decision list

RoadOrientation = StraightPlain AND
RoadJunction = Roundabout AND
Subcity = Kolfe: Fatal (4.0/0.0)

RoadOrientation = StraightPlain AND
Subcity = Arada AND
RoadSeparation = BiDirectional: PropertyLoss (52.19/18.01)

RoadOrientation = StraightPlain AND
Subcity = Arada: SevereInjury (10.06/5.06)

RoadOrientation = StraightPlain AND
ParticularArea = MarketArea: SevereInjury (4.04/0.04)

Subcity = Kirkos AND
RoadJunction = T-Shape AND
ParticularArea = Office: PropertyLoss (974.83/81.11)

Subcity = Kirkos AND
RoadJunction = CrossRoad AND
ParticularArea = Office AND
RoadSeparation = Island: PropertyLoss (410.83/32.04)

Subcity = Gulele AND
RoadSeparation = BiDirectional: SevereInjury (45.1/20.05)

Subcity = Lafto AND
ParticularArea = Churches AND
RoadSeparation = BiDirectional: SevereInjury (16.27/4.08)

Figure 1: Partial output from the PART rule generator.

The rules above indicate that accident severity varied with different combinations of road-related factors. For instance, there were more scenarios for severe injury on straight plain roads than other orientations of roads in the same sub-city.

Significance and Contribution of the Study

The significance of this research lays in its development of new insights related to road accidents in Ethiopia. These insights will provide valuable help in developing methods to improve road safety, particularly in the phase of choosing appropriate means and budget allocations of resources. Considering the size of the accident data set, applying data mining techniques to model RTA data records can help to reveal how the drivers' behavior and roadway and weather conditions are causally connected with different injury severities. This can help decision makers to formulate better traffic safety control policies, label roads with necessary signs informing drivers and pedestrians of accident risks, and design better roads. Another expected outcome of the research is a better understanding of the suitability of data mining methods to the safety research field.

Conclusion

A thorough literature review revealed a gap in published studies on the relationship between road characteristics and RTA severity in Ethiopia. In this paper, we collected and cleaned traffic accident data, attempted to construct novel attributes, and tested a number of predictive models. The outputs of the models were presented for analysis to domain experts for feedback. The RTA is eager to continue the study to identify areas of interest that should be given resources for traffic safety. Finally, knowledge was presented in the form of rules using the PART algorithm of WEKA.

In contrast with the previously published work of the authors, which focused on driver characteristics, here we focused on the contribution that various road-related factors have on the accident severity. The results of this study could be used by the respective stakeholders to promote road safety. While the methods are simple, the results of this work could have tremendous impact on the well-being of Ethiopian civilians. The next step in the modeling will be to combine road-related factors with driver information for better predictions, and to find interactions between the different attributes. We also plan to develop a decision support tool for the Ethiopian Traffic Office.

References

Abdalla, I. M., R. Robert, et al. (1997). "An investigation into the relationships between area social characteristics and road accident casualties." Accidents Analysis andPreventions 5: 583-593.

Beshah, T. (2005). Application of data mining technology to support RTA severity analysis at Addis Ababa traffic office. Addis Ababa, Addis Ababa University.

Beshah, T., A. Abraham, et al. (2005). "Rule Mining and Classification Of Road Traffic Accidents Using Adaptive Regression Trees." Journal Of Simulation 6(10-11).

Chang, L. and W. Chen (2005). "Data mining of tree-based models to analyze freeway accident frequency." Journal of Safety Research 36: 365-375.

Chang, L. and H. Wang (2006). "Analysis of traffic injury severity: An application of non-parametric classification tree techniques Accident analysis and prevention " Accident analysis and prevention 38(5): 1019-1027.

Chong, M., A. A., et al. (2005). "Traffic Accident Analysis Using Machine learning Paradigms." Informatica 29(1).

Getnet, M. (2009). Applying data mining with decision tree and rule induction techniques to identify determinant factors of drivers and vehicles in support of reducing and controlling road traffic accidents: the case of Addis Ababa city. Addis Ababa Addis Ababa University.

Mussone, L., A. Ferrari, et al. (1999). "An analysis of urban collisions using an artificial intelligence model." Accident Analysis andPrevention 31: 705-718.

Ng, K. S., W. T. Hung, et al. (2002). "An algorithm for assessing the risk of traffic accidents." Journal of Safety Research 33: 387-410.

Ossenbruggen, P. J., J. Pendharkar, et al. (2001). "Roadway safety in rural and small urbanized areas." Accidents Analysis and Prevention 33(4): 485-498.

Sohn, S. and S. Hyungwon (2001). "Pattern recognition for a road traffic accident severity in Korea." Ergonomics 44(1): 101-117.

Sohn, S. and S. Lee (2002). "Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. ." Safety Science 41(1): 1-14.

Srisuriyachai, S. (2007). Analysis of road traffic accidents in Nakhon Pathom province of Bangkok using data mining. Graduate Studies. Bangkok, Mahidol University.

Witten, I. H. and E. Frank (2000). Data Mining: practical machine learning tools and techniques with java implementations. San-Francisco, Morgan Kaufmann publishers.

Wong , J. and Y. Chung (2008). "Comparison of Methodology Approach to Identify Causal Factors of Accident Severity." Transportation Research Record 2083: 190-198.

Zelalem, R. (2009). Determining the degree of driver’s responsibility for car accident: the case of Addis Ababa traffic office. Addis Ababa, Addis Ababa University.