Traffic Accident Analysis from Drivers Training Perspective Using Predictive Approach

Hiwot Teshome
hewienat@gmail

Tibebe Beshah
School of Information Science, Addis Ababa University, Ethiopia
tibebe.beshah@gmail.com

Abstract

Road traffic accidents are among the top leading causes of deaths and injuries, especially in the developing world. Ethiopia in particular experiences the highest rate of such accidents. One of the solutions to reduce the problem of traffic accident is finding the causes through research studies. In this work, factors for the relationship between driver's training and road traffic accident are identified using data mining classification technique. Using WEKA as a tool, different rules are generated so that different domain experts get new insights related to road traffic accidents.

Keywords: Accident; Data mining; Classification; WEKA

1. Introduction

Traffic accident is a special case of trauma that constitutes a major cause of disability and untimely death [1]. As reported by WHO [2], around 1.24 million people die each year as a result of road traffic crashes. Without action, road traffic crashes are predicted to result in the deaths of around 1.9 million people annually by 2020.

Traffic accident is the result of multiplicity of factors and it is often the interaction of more than one variable that leads to the occurrence of accident among which are driver, road and traffic characteristics [3]. For the last five years records have shown that over 87% of all road accidents in Ethiopia are attributed to driver error [4].

Driver factors in road traffic accidents are all factors related to drivers and other road users. This may include driver behavior, visual and auditory acuity, decision making ability and reaction speed. Drug and alcohol use while driving is an obvious predictor of road traffic accident, road traffic injury and death. Speeding, travelling too fast for prevailing conditions or above the speed limit, is also a driver factor that contributes to road traffic accidents [5].

The capital city of Ethiopia, Addis Ababa, shares 65% of the total accident in the country. Pedestrians are the most vulnerable ones in Addis Ababa; over 81% of accident fatalities are of accident type "car hit pedestrian". Moreover, 81% of crashes in Ethiopia are attributed to driver error [6].

In an attempt to prevent road accidents one role that can be played is finding the factors through research. Researches are done using different tools and technologies. In line with this, data mining is one of these research technologies which is growing rapidly through time [7].

In order to combat this problem, various road safety strategies have been proposed and used. Research on road safety has been conducted for several years, yet issues like drivers training, driver experience still remain undisclosed and unresolved. Besides that, research on road safety using data mining tools has been conducted for several years also. The aim of this research is identifying the relationship between drivers training and road traffic accident using classification technique of data mining.

The previous researches so far, directly or indirectly, try to identify the factors of road traffic accidents on the side of road, drivers and so on but even if the main factor fall on the shoulder of the driver as per the author's knowledge from the training perspective of drivers is not considered yet. If that is the case mining from the driver's training point of view is worthwhile.

Through the research attempt has been made to answer the following research questions:

- ✓ What are the determinant attributes from driver's training related factors that have impact on the occurrence of traffic accident?
- ✓ What interesting patterns or rules can be generated?
- ✓ Which data mining technique performs well in developing a model that can explain the relationship between drivers training and road traffic accident?

The primary data were gathered using questioners. 1000 questioners were distributed for different drivers that have different status, knowledge, occupation, age and gender. To have a better understanding of the problem the sampling was taken from different associations that work on drivers like taxi drivers association groups, heavy truck drivers' association and the data available from Ethiopian Road Authority. As per the author's knowledge, since this kind of research was not done before literatures were not used as a baseline to develop a questionnaire. Because of that a questionnaire was developed using the data that was analyzed during data understanding and domain understanding and also with the help of experts who work in the driver training area. As a pilot test 20 questionnaires were distributed for 20 drivers; 10 males and 10 females with the average age from 20-45 and their feedback was collected and helped the questionnaire to be refined again.

To prepare the data collected from the questionnaire into a suitable form for data mining analysis, data preparation activities like attribute selection, data cleaning, and data formatting have been done. After automating the data on Excel sheet, the first task was selecting attributes that have direct relationship with the objective. After the feature selection of the data using WEKA tool, the preprocessing stage which includes cleaning inconsistent and incorrect data, incomplete records, predict missing values, correct erroneous and anomalous data is performed. After the pre processing steps the data contain the selected predictors in a format which is ready to use for modeling.

In this research classification of accident occurrence is done attempting different classification techniques like J48, PART and Naïve Bayes to predict the occurrence of accident. Among the tools WEKA software is used to conduct the research.

2. Related Work

Many data mining application researches on the analysis of road traffic accident have been done globally and a few locally. Reviewing this related works that were done using data mining tools and techniques at different place and time in the same problem domain gave the author an in-depth insight for this research.

As shown by Tibebe [8], data mining is one technique in order to conduct a research on historical road traffic accidents to investigate analysis accident severity where the data comprising a dataset of 4,658 accident records at Addis Ababa Traffic Office. In this work various classification models using the decision tree technique by applying Knowledge SEEKER algorithm of the Knowledge STUDIO data mining tool were built to help in decision-making process at the traffic office. The methodology adopted had three basic steps namely data collection, data preparation, and model building and validation.

The model classifies accident severity into four classes as 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 identified as the basic determinant variables for injury severity level. The classification accuracy of the decision tree classifier was found to be 87.47%.

Zelalem [9] conducted a data mining research to classify driver's responsibility on a given accident in

Addis Ababa. The research uses decision tree and multilayer perception (MLP) neural network data mining techniques to analyze the accident data. The study focuses on predicting the degree of driver's responsibility for car accidents and identifying the important factors influencing the different levels of responsibility using the road traffic accident dataset of Addis Ababa Traffic Control and Investigation Department (AARTCID).

The researcher used WEKA data mining tool to build the decision tree (using the ID3 and J48 algorithms) and MLP (the back propagation algorithm) predictive models. Rules representing patterns in accident dataset have been extracted from the decision tree indicating important relationships between variables that influence driver's degree of responsibility such as age, license grade, level of education, driving experience, and other environmental factors. According to the author, the accuracies of the models were 88.24% and 91.84% respectively. In addition the research reveals that the decision tree model is found to be more appropriate for the problem under consideration.

On the other hand Getnet [10] 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 researcher used WEKA 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.

In this research determinant factors of drivers and road that cause traffic accident are identified which will help health policy makers in planning health programs and it will also support the Traffic Control Division of Addis Ababa in taking proper action such as revising the existing traffic rules against vehicle accidents.

As pointed out by Tibebe and Hill [11], developing adaptive regression trees to build a decision support system is one of the methods to handle road traffic accident. The study focused on injury severity levels resulting from an accident using real data obtained from Addis Ababa traffic office.

Yang *et al.* [12] used Neural Network approach to detect safer driving patterns that have less chances of causing death and injury when a car crash occurs. They performed the Cramer's "V" Coefficient test to identify significant variables that cause injury, therefore, reduced the dimensions of the data for the analysis.

Sohn and Lee [13] applied data fusion, ensemble and clustering to improve the accuracy of individual classifiers for two categories of severity (bodily injury and property damage) of road traffic accident. They applied a clustering algorithm to the dataset to divide the data into subsets of data, and then used each subset of data to train the classifiers (Neural Network and decision trees).

Table1: Summary of Researches

Author	*Objective*	*Method*	*Key Result*
Tibebe (2005)	To investigate analysis accident severity	decision tree technique	Accident cause, accident type, road condition, vehicle type, light condition, road surface type and driver age as the basic determinant variables for injury severity level.
Zelalem (2009)	To classify driver's responsibility	decision tree and multilayer perception (MLP) neural network	Age, license grade, level of education, driving experience influence driver's degree of responsibility

Author	*Objective*	*Method*	*Key Result*
Getnet (2009)	To identify and predict major driver and vehicle risk factors that cause RTAs	decision tree	The license grade, vehicle service year, vehicle type, and experience are most important variables for predicting accident severity.
Tibebe & Hill (2010)	Mining Road Traffic Accident Data to Improve Safety	adaptive regression tree	road-related factors are identified

As shown in Table 1 different scholars did their researches on road traffic accident using data mining techniques and they have proposed their solutions that they identified as factors for the occurrence of an accident. On the contrary, this research tries to identify factors of traffic accident with respect to the training view of drivers; that is to mean those attributes that are generated as factors on drivers training instead of being general factors for road traffic accident.

3. The Proposed Solution

In addressing the above mentioned problem the data mining experiment process and the resulted prototype are presented in this section.

The underlying problem that initiated this project is the fact that road traffic accidents are among the top leading causes of death and injury of various levels in Ethiopia. During the document analysis it was found that the data source for this research could not be the traffic accident data kept at AARTCID because majority of the data that is collected from did not correspond to the research objective. In order to come up to the solution the first thing that has been done was preparing and distributing questionnaire to use it as a data set for this research.

As can be found from data of Ethiopian Road Transport Authority the total number of vehicles that are found in all regions of Ethiopia are 474,143 and in Addis Ababa in particular is 219,217.

From all types of vehicles, public transport, automobile, trailers dry (*Derek chenet*) and trailer liquids (*fesash chenet*) have been selected with their respective values of 48%, 32%, 12%, and 8% which is taken as the total population for this research.

Using proportional stratified sampling technique 1000 samples were taken to distribute the questionnaire for the four categories of vehicle type as shown in Table 2.

Table 2: Number of Distributed questionnaires

Type of car	*Number of questionnaires distributed*
Automobile	300
Public Transport	500
Derek chenet	100
Fesash chenet	100

After the distribution and collection of questionnaire the second task was encoding it into an Excel sheet so that it will get prepared for data the cleaning step. The encoded data set that is generated from the questionnaire have 17 columns (or attributes) of text and numbers. On the data cleaning steps noises, missing values, and redundant values were treated. Finally 10 attributes remained through feature selection for the modeling purpose as shown in Table 3.

Table 3: Final Selected attributes with their descriptions

No.	*Attribute Name*	*Description*
1.	DriverAge	Age
2.	DriverSex	Sex
3.	DriverEducational status	Educational level of the driver
4.	DriverYear of licence	The time to get the license
5.	DriverCartype	Category of the driver's car
6.	DriverExperience	Driving experience
7.	OccurredAccident	Occurred accident
8.	DriverBesttraning	Training that the driver chooses
9.	Type of traningForBehaviour Change	Type of the training which cause behavioral change
10.	TrainingTypeToCauseInfluence	Type of training to influence accident

During modeling several experiments are made using the J48, PART and Naïve Bayes to come up with a meaningful output with two test options. Major factors for the relationship between accident and training are identified and rules are generated using J48 decision trees and rule induction (PART algorithm). The comparison of the models using WEKA's experimenter showed that PART slightly outperforms Navies and J48 algorithms with an accuracy of 92.4%, 91.2% and 90.9% respectively. The results of the experiments carried out in this research show driver age, driver educational status, and car type are determinant factors for drivers training which can help to predict accident occurrence.

A prototype is developed using rules from PART and J48 algorithms. This helps to have a better understanding from the two rules generated. From both algorithms different rules were generated and the rule which can determine driver related factors can be shown as fines and new rules. Making the prototype to predict differently as per the algorithm helps to analyze the output in different perspectives.

4. Discussion

Three experiments are conducted using 10 attributes that are filtered during the data preparation phase with the selected algorithms (J48, PART, Navies). Based on the selected algorithms different experiments have been done using two test options (tenfold cross validation 66% percentage split) and different results were obtained.

In the series of experiments, evaluation of models is done based on accuracy of models and confusion matrix. All the three classifiers performed well and almost similarly with respect to the number of correctly classified instances. Evaluating the models helps to identify a relatively better model. Table 4 shows the evaluations of the three algorithms compared with respect to these efficient checking algorithms.

Table 4: Summary of the result of the three experiments

No.	*Classification Models(classifiers)*	*Number of Correctly classified instances*		*Accuracy in Percentage*	
		10 fold cross validation	*66% percentage split*	*10 fold cross validation*	*66% percentage split*
1	J48	912	304	91.2%	89.4118
2	PART	924	304	92.4%	89.4118
3	Naïve Bayes	909	303	90.9%	89.1176

As mentioned earlier we can clearly see from Tables 5 and 6 that there is a relative better model prediction in the case of PART incorrectly identifying the dataset. The ROC area of the PART indicates 0.93265 with small significance difference with ROC area in Naïve Bayes indicates 0.9334 and also much better than the J48 ROC area. The overall model accuracy of PART is 92.4% which shows it has better prediction than the accuracy of J48 and Naïve Bayes. For the given data set under this study PART with 10 fold cross validation has shown better accuracy.

Table 5: Confusion matrix Predicted category

Actual Category	*Predicted category*		
	J48	*PART*	*Naïve Bayes*
Yes	782	783	758
No	66	55	45

Once the model has been trained and tested, we need to measure the performance of the model. In data mining comparison is done by using accuracy and ROC (Receiver Operating Characteristic).

Table 6: Comparison of Algorithms using ROC

Algorithm Efficiency Evaluator	*J48*	*PART*	*Naïve Bayes*
ROC	0.8878	0.9365	0.9334

A prototype is developed using rules from PART and J48 algorithms. This helps to have a better understanding from the two rules generated. From both algorithms different rules were generated and the rule which can determine driver related factors were fines and new rules. Making the prototype to predict differently as per the algorithm helps to analyze the output in different perspectives.

5. Conclusion and Future Work

This research attempted to study the possible application of classification techniques of data mining to predict the relationship between drivers training and road traffic accident. The study followed CRISP DM model and WEKA data mining tool has been used to implement the Naïve Bayes, J48, and PART algorithms.

The results of the experiments carried out in this research show driver age, driver educational status, car type are determinant factors for drivers training which can help to predict accident occurrence. A prototype is developed using rules from PART and J48 algorithms. This helps to have a better understanding from the two rules generated. From both algorithms different rules were generated and the rule which can determine driver related factors were fines and new rules. Making the prototype to predict differently as per the algorithm helps to analyze the output in different perspectives.

In general, encouraging results are obtained by employing both decision tree and rule induction techniques and the rules generated by J48 and PART algorithms are easily understandable by subject experts. Thus, the results obtained in this research have proved the applicability of data mining in road traffic accident preventing and controlling activities especially on the area of drivers training. More specifically it will provide valuable help in developing new methods, rules and regulations to increase the quality of the drives training now a day.

This work focused only on drivers training using classification techniques. The next step will be applying different techniques like clustering and associating data mining techniques integrating with driver information for better predictions, and to find interactions between the different attributes.

References

[1] Dipo T. Akomolafe Akinbola Olutayo, “Using Data Mining Technique to Predict Cause of Accident and Accident Prone Locations on Highways”, American Journal of Database Theory and Application 2012, 1(3): 26-38.

[2] WHO, World report on road traffic injury prevention, Switzerland, Geneva, 2013.

[3] William Eckersley, Ruth Salmon, and Mulugeta Gebru, “Khat, Driver Impairment and Road Traffic Injuries: A View from Ethiopia”, Jerusalem Children and Community Development Organization, Addis Ababa, Ethiopia, January 2010.

[4] Ethiopian Road Authority, "How safe are Ethiopian roads?", Unpublished road safety report, Addis Ababa, Ethiopia, 2005.

[5] Road Traffic Accidents in Nigeria: A Public Health Problem, 2013.

[6] A. Persson, "Road traffic accidents in Ethiopia: magnitude, causes and possible interventions", April 2008.

[7] The CRISP-DM consortium, "Step-by-step Data Mining Guide Available", August, 2000.

[8] Tibebe Beshah, "Application of data mining technology to support RTA severity analysis at Addis Ababa Traffic Office", Unpublished Master's Thesis, Addis Ababa University, 2005.

[9] Zelalem Regassa, "Determining the degree of drivers' responsibility for car accident: the case of Addis Ababa traffic office", Unpublished Master's Thesis, Addis Ababa University, 2009.

[10] Getnet M., "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", Unpublished Master's Thesis, Addis Ababa University, 2009.

[11] Tibebe Beshah and Shawndra Hill, "Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia", 2010. Available at www. ai-d.org/pdfs/Beshah.pdf, Last accessed on March 13, 2013.

[12] Yang, W.T., Chen, H. C., and Brown, D. B., "Detecting Safer Driving Patterns by A Neural Network Approach", ANNIE '99 for the Proceedings of Smart Engineering System Design Neural Network, Evolutionary Programming, Complex Systems and Data Mining, Vol. 9, pp. 839-844, Nov. 1999.

[13] Sohn, S. Y. and Lee, S. H., "Data Fusion, Ensemble and Clustering to Improve the Classification Accuracy for the Severity of Road Traffic Accidents in Korea", Safety Science, Vol. 4, Issue1, pp. 1-14, February 2003.