I.J. Information Technology and Computer Science, **2014, 02, 22-28**
Published Online January 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2014.02.03

Traffic Accident Analysis Using Decision Trees and Neural Networks

Olutayo V.A
Computer Science Department, Joseph Ayo Babalola University, Ikeji-Arakeji, Nigeria
Corresponding E-mail: vicsy2004@yahoo.com

Eludire A.A
Computer Science Department, Joseph Ayo Babalola University, Ikeji-Arakeji, Nigeria
E-mail: aaeludire@yahoo.com

Abstract— This work employed Artificial Neural Networks and Decision Trees data analysis techniques to discover new knowledge from historical data about accidents in one of Nigeria's busiest roads in order to reduce carnage on our highways. Data of accidents records on the first 40 kilometres from Ibadan to Lagos were collected from Nigeria Road Safety Corps. The data were organized into continuous and categorical data. The continuous data were analysed using Artificial Neural Networks technique and the categorical data were also analysed using Decision Trees technique .Sensitivity analysis was performed and irrelevant inputs were eliminated. The performance measures used to determine the performance of the techniques include Mean Absolute Error (MAE), Confusion Matrix, Accuracy Rate, True Positive, False Positive and Percentage correctly classified instances. Experimental results reveal that, between the machines learning paradigms considered, Decision Tree approach outperformed the Artificial Neural Network with a lower error rate and higher accuracy rate. Our research analysis also shows that, the three most important causes of accident are Tyre burst, loss of control and over speeding.

Index Terms— Traffic Accident Data Mining, Accident Causes Prediction and Sensitivity Analysis, Performance Comparison

I. Introduction

The problem of deaths and injuries as a result of accidents is acknowledged to be a global phenomenon and traffic safety has been a serious concern since the start of the automobile age, almost one hundred years ago. It has been estimated that over 300,000 persons die and 10 to 15 million persons are injured every year in road accidents throughout the world. Statistics have also shown that mortality in road accidents is very high among young adults that constitute the major part of the work force. In order to combat this problem, various road safety strategies, methods and counter measures have been proposed and used. These methods mainly involve conscious planning, design and operations of roads. One important feature of this method is the identification and treatment of accident-prone locations commonly called black spots [1].

However, black spots are not the only causes of accidents on the highway. Regression analysis is a common approach used in modelling highway geometrics, traffic characterizations and accident frequencies in order to determine other causes of accidents. Regression analysis highly depends on traffic flow data such as Average Daily Traffic (ADT). It also requires the researcher to know exactly the dependent variables as well as the independent variables. Sadly however, in Nigeria, data are often looked at from one dimension. More often than not, the causes for road accidents in developing country like Nigeria may have nothing to do with the highway geometry, or even traffic characterization.

Also, a large number of data mining algorithmic solution exist; but until now, little or no empirical research has been done on comparing their efficiency especially on road accidents data set. Therefore, this research work is useful to ascertain which of these data mining classification's algorithmic solutions will scale better (Artificial Neural Networks and Decision Trees) on road accident database.

Finally, the purpose of this research is to look at historical data of road accidents on one of the Nigeria's busiest roads on how can be more analysed in order to discover new knowledge about road accidents in Nigeria and use this knowledge to reduce the carnage on our high way.

Related important works can be summarized as follows. Abdelwahab et al. [2] studied the 1997 accident data for the Central Florida area focusing on two-vehicle accidents that occurred at signalized intersections. The injury severity was divided into three classes: no injury, possible injury and disabling injury. The performance of Neural Network (NN) trained by Levenberg-Marquardt algorithm and Fuzzy ARTMAP were compared, and found that NN (65.6% and 60.4%

classification accuracy for the training and testing phases) performed better than Fuzzy ARTMAP (56.1%). Bedard et al. [3] applied a multivariate logistic regression to determine the independent contribution of driver, crash, and vehicle characteristics to drivers' fatality risk. It was found that increasing seatbelt use, reducing speed, and reducing the number and severity of driver-side impacts might prevent fatalities. Some researchers studied the relationship between drivers' age, gender, vehicle mass, impact speed or driving speed measure with fatalities [4,] Dia et al. used real-world data for developing a multi-layered NN freeway incident detection model [5]. Results showed that NN could provide faster and more reliable incident detection over the model that was in operation on Melbourne's freeways.

Evanco conducted a multivariate population-based statistical analysis to determine the relationship between fatalities and accident notification times [6]. The analysis demonstrated that accident notification time is an important determinant of the number of fatalities for accidents on rural roadways. Kim et al. [7, 8] developed a log-linear model to clarify the role of driver characteristics and behaviours in the causal sequence leading to more severe injuries. They found that driver behaviours of alcohol or drug use and lack of seat belt use greatly increase the odds of more severe crashes and injuries.

Shankar, et al. [9] applied a nested logic formulation for estimating accident severity likelihood conditioned on the occurrence of an accident. The study found that there is a greater probability of evident injury or disabling injury/fatality relative to no evident injury if at least one driver did not use a restraint system at the time of the accident. Ossiander et al. [10,11,12,13] used Poisson regression to analyse the association between the fatal crash rate (fatal crashes per vehicle mile travelled) and the speed limit increase and found that the speed limit increase was associated with a higher fatal crash rate and more deaths on freeways in Washington State.

Yang, et al. used NN approach to detect safer driving patterns that have less chances of causing death and injury when a car crash occurs [14,15,16]. They performed the Cramer's V Coefficient test [17, 18] to identify significant variables that cause injury, therefore, reduced the dimensions of the data for the analysis. The 1997 Alabama interstate alcohol-related data was used and was found that by controlling a single variable (such as driving speed, or light conditions) fatalities and injuries could be reduced by up to 40%.

Akomolafe O.P. [19] employed Artificial Neural Network using Multilayer perceptron to predict likelihood of accident happening at particular location between the first 40 kilometres along Lagos-Ibadan Express road. He used Neurosolution version 4.1 software from Neurodimension Inc. on a Pentium III. He performed sensitivity analysis to extract redundant input and discovered that location 2 recorded the highest number of road accident occurrence and found out that, tyre burst was the major cause of accident along the route after implementing training, verification and test data set.

The remaining part of this paper is organized as follows: In Section II, we discuss the accident data set used in our work. Section III is devoted to introduction of the machine learning techniques utilized in our work, performance results and their analysis are also presented here.

II. Accident Data Set

The heart of any decision tree and artificial neural network prediction model is relevant and historical data of the domain in consideration. The selection of inputs is the most important aspect of creating a useful prediction, as it represents all of the knowledge that is available to the model to base the prediction on. This study used dataset from the Nigeria Road Safety Corps. The data sample used in this study covered a period of twenty four Months from January 2002 to December 2003 on the first 40 kilometres from Ibadan to Lagos. It consists of label-variables as given in table 1.

Table 1: Continuous and Categorical Variables values

S/N	Variable	Description	Value	Type
1.	Vehicle Type	Small cars	1	Categorical
		Heavy Vehicle	2	Categorical
2.	Time of the day	Morning	1	Categorical
		Afternoon	2	Categorical
		Evening	3	Categorical
		Night / Midnight	4	Categorical
3.	Season	Wet	1	Categorical
		Dry	2	Categorical
4.	Causes	Wrong Overtaking	A	Categorical
		Careless Driving	B	Categorical
		Loss of Control	C	Categorical
		Tyre Bust	D	Categorical
		Over Speeding	E	Categorical
		Obstruction	F	Categorical
		Pushed by another vehicle	G	Categorical
		Broken Shaft	H	Categorical
		Broken Spring	I	Categorical
		Brake Failure	J	Categorical
		Road problem	K	Categorical
		Unknown Causes	L	Categorical
		Robbery Attack	M	Categorical

The unknown causes may include other factors such as;

i. Law enforcement agents problems
ii. Driver's condition
iii. Attitude of other road users
iv. Inadequate traffic road signs
v. Condition of the road surface
vi. Demographic factors of the location of accident
vii. Traffic congestion
viii. Vehicle make
ix. Vehicle year of manufacturing
x. General Vehicle conditions

The output variable is the location, critical study of the accident data showed that the locations can be divided into three distinct regions tagged region A, region B and region C, meaning we have three outputs. Where,

First location 1 – 10km is Region A or location 1

Between 10km – 20km is region B or Location 2

Above 20km is region C or Location 3

In this research, the datasets were organized into both categorical data and continuous data. The decision tree was given the categorical data and the artificial neural networks was equally given continuous data, this is to allow the data mining algorithms to run on different data type, so as to determine which performs better. The major step required to obtain result of the research involved analysis of the data using WEKA. WEKA is a collection of machine learning algorithms and data processing tools. It contains various tools for data pre-processing, classification, regression, clustering, association rules and visualization. There are many learning algorithms implemented in WEKA including Bayesian classifier, Trees, Rules, Functions, Lazy classifiers and miscellaneous classifiers. The algorithms can be applied directly to a data set. WEKA is also data mining software developed in JAVA it has a GUI chooser from which any one of the four major WEKA applications can be selected. For the purpose of this study, the Explorer application was used.

III. Experimental Setup, Analysis and Results

3.1. Artificial Neural Networks Analysis

In the case of ANN based modelling, the hyperbolic activation function was used in the hidden layer and the logistic activation function otherwise known as sigmoid in the output layer. Models were trained with BP (100 epochs, learning rate 0.01) and SCGA (500 epochs) to minimize the root mean square and mean absolute error. For each output class, both multilayer perception (MLP) and Radical Basis Function Neural Networks (RBF) were used to determine the better networks.

3.2. Radial Basis Function Performance Analysis

The RBF model was experimented with using different number of hidden neurons, and the model with highest classification accuracy for the correctly classified instances was determined. From the result analysis, the RBF model achieved training and testing performance of 54.73% and 40.56% respectively with 0.3478 of mean absolute error.

Table 2: Detailed Accuracy by Class

Class	Roc Area	TP rate	FT rate	Precision	Recall	F- measure
Location (3)	0.716	0.294	0.096	0.476	0.294	0.364
Location (2)	0.517	0.744	0.694	0.598	0.744	0.663
Location (1)	0.568	0.25	0.108	0.35	0.25	0.292
Weighted Avg.	0.572	0.547	0.446	0.523	0.547	0.524

Table 3: Confusion Matrix

	Predicted Category		
Actual Category	Location (3)	Location (2)	Location (1)
Location (3)	10	23	1
Location (2)	10	64	12
Location (1)	1	20	7

3.3. Multilayer Perception Performance Analysis

For the case of MLP model, the model achieved training and testing performance of 78 correctly classified instances representing 52.70% and 28 representing 45.20% with mean absolute error of 0.3479 and root mean square error of 0.5004.

Table 4: Detailed Accuracy by Class

Class	Roc Area	TP rate	FT rate	Precision	Recall	F- measure
Location (3)	0.529	0.158	0.5	0.529	0.514	0.719
Location (2)	0.628	0.581	0.6	0.628	0.614	0.493
Location (1)	0.214	0.133	0.273	0.214	0.24	0.564
Weighted Average	0.527	0.399	0.515	0.527	0.52	0.558

Table 5: Confusion Matrix

	Predicted Category		
Actual Category	Location (3)	Location (2)	Location (1)
Location (3)	18	15	1
Location (2)	17	54	15
Location (1)	1	21	6

A confusion matrix provides detailed information about how data rows are classified by the model. The matrix has a row and column for each category of the target variable (Location).

The categories shown in the first Column are the actual categories of the target variable.

The categories shown across the top of the table 5 cells are the predicted categories. The numbers in the cells are weights of the data rows with the actual category of the row and the predicted category of the column.

The numbers in the diagonal cells are the weights for the correctly classified cases where the actual category matches the predicted category the off-diagonal cells have misclassified row weights.

For both RBF and MLP, the confusion matrix showed that the model gave a good performance on location2. Further investigation into the input data especially on the importance of variables revealed that Tyre burst has the highest value of all the sixteen variables, followed by loss of control and over speeding.

3.4. Sensitivity and Specificity Report

The sensitivity and specificity report is used for classification problems where the target variable has two categories. For these types of analyses, one category of the target variable is called the 'positive' category, and the other is called the 'negative' category. These are the parameters used to measure the performance and the accuracy rate of the models. That is,

TP-represent True positive

FP-represent false positive and

ROC- represent receive operating characteristic curve for the model

ROC is also called the "C statistic".

3.5. Decision Tree Performance Analysis

Several numbers of setups of decision tree algorithms have been experimented and r the best results obtained is reported for my data set.

Each class was trained with entropy of fit measure, the prior class probabilities parameter was set to equal, the stopping option for pruning was misclassification error, the minimum n per node was set to 5, the fraction of objects was 0.05, the maximum number of nodes was 100, surrogates was 5, 10 fold cross-validation was used, and generated comprehensive results.

The best decision tree result was obtained with Id3 with 115 correctly classified instances and 33 incorrectly classified instances which represents 77.70% and 22.29% respectively.

Mean absolute error was 0.1835 and Root mean squared error was 0.3029.

3.5.1. Decision Tree Performance Analysis on Id3

Table 6: Detailed Accuracy by Class

Class	TP rate	FT rate	Precision	Recall	F- measure	Roc Area
Location (3)	0.688	0.069	0.733	0.688	0.71	0.942
Location (2)	0.897	0.361	0.78	0.897	0.834	0.888
Location (1)	0.517	0.025	0.833	0.517	0.638	0.95
Weighted Average	0.777	0.232	0.78	0.777	0.769	0.912

Table 7: Confusion Matrix

	Predicted Category		
Actual Category	Location (3)	Location (2)	Location (1)
Location (3)	22	10	0
Location (2)	6	78	3
Location (1)	2	12	15

3.5.2. Decision Tree performance Analysis on Function Tree (FT)

Table 8: Detailed Accuracy by Class

Class	TP rate	FT rate	Precision	Recall	F- measure	Roc Area
Location (3)	0.625	0.086	0.667	0.625	0.645	0.869
Location (2)	0.77	0.361	0.753	0.77	0.761	0.736
Location (1)	0.586	0.101	0.586	0.586	0.586	0.832
Weighted Average	0.703	0.25	0.702	0.703	0.702	0.783

Table 9: Confusion Matrix

	Predicted Category		
Actual Category	Location (3)	Location (2)	Location (1)
Location (3)	20	12	0
Location (2)	8	67	12
Location (1)	2	10	17

3.6. Discussion

In the case of neural networks based modelling, two types of algorithms were used: Multilayer perceptron MLP and Radial Basis Function (RBF).

Models were trained with 500 epochs to minimize the root mean square and mean absolute error.

For the RBF model, different numbers of hidden neurons were experimented and report the model with highest classification accuracy for the correctly classified instances. From the result, RBF model achieved training and testing performance of 54.73% and 40.56% respectively with 0.3478 of mean absolute error and 0.4484 of root mean square error. Also from the detailed accuracy by class and from confusion matrix of the result, RBF attained an accuracy rate of 0.547.

For the MLP model, the model achieved training and testing performance of 78 correctly classified instances representing 52.70% and 28 incorrectly classified instances representing 45.20% with mean absolute error of 0.3479 and mean square error of 0.5004. From the detailed accuracy by class, MLP attained an accuracy rate of 0.399.

In the case of Decision Tree Performance analysis, the, dataset were experimented with two algorithms. They are Id3 and FT (function tree)

For Id3 algorithm, there are 115 correctly classified instances and 33 incorrectly classified instances which represent 77.70% and 22.29% respectively. Mean absolute error was 0.1835 and Root mean squared error was 0.3029.

Also for functional tree algorithm (FT), total number of tree size was 5 with 105 correctly classified instances representing 70.27% and 44 incorrectly classified instances representing 29.73%.

From the detailed accuracy by class and confusion matrix, Id3 attained accuracy rate of 0.777 and FT attained accuracy rate of 0.703.

Finally, comparing the techniques from the result analysis shows that Decision Tree performs better than the Neural Networks based on the error report, number of correctly classified instances and accuracy rate generated.

Table 10: Summary of Performance

Performance Measure	Neural Networks	Decision Tree
Mean absolute error rate	**RBF Networks** 0.3478	**Id3** 0.1835
Correctly classified instances %	54.73	77.70
Accuracy rate	0.547	0.777
Mean absolute error rate	**MLP Networks** 0.3479	**FT** 0.2519
Correctly classified instances %	52.70	70.27
Accuracy rate	0. 399	0.703

On the attribute selection, using chi-squared attribute evolution with ranking method, tyre burst which represent attribute 7 has the highest value of 13.7826 followed by broken-shaft with 11.1 and loss of control with 10.8756 Also, Location two has the highest record of accidents with tyre burst being the major cause of accident on the highway.

IV. Conclusion

In this paper, a comparison of different Decision Tree algorithms and Artificial Neural Networks performance, were analysed on road accident data set. The location is between the first 40 kilometres along the Ibadan-Lagos Express road. The work used Multilayer Perceptron as well as Radial Basis Function (RBF) Neural Networks and Id3 and Function Tree algorithms.

Results shows that the Id3 tree algorithm performed better with higher accuracy rate, while Radial basis function performed better than multilayer perceptron in terms of time used in the building of the model and number of correctly classified instances. Finally, our experiments showed that, Decision Tree techniques outperformed Artificial Neural Networks with a lower error report and with a higher number of correctly classified instances and better accuracy rate generated. Tyre burst, broken shaft and loss of control variables were the three major causes of accidents where tyre burst represents the major cause of accidents.

References

[1] Olutayo, V.A. (2011); Comparison of different data mining techniques performance in knowledge discovery from road accident database. M.Sc. Thesis, Department of computer science, University of Ibadan, Nigeria.

[2] Abdelwahab, H. T. & Abdel-Aty, M. A. Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. *Transportation Research Record 1746*, Paper No. 01-2234.

[3] Bedard, M., Guyatt, G. H., Stones, M. J., & Hireds, J. P., The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities. *Accident analysis and Prevention*, Vol. 34, 2002, pp. 717-727.

[4] Buzeman, D. G., Viano, D. C., & Lovsund, P., Car Occupant Safety in Frontal Crashes: A Parameter Study of Vehicle Mass, Impact Speed, and Inherent Vehicle Protection. *Accident Analysis and Prevention*, Vol. 30, No. 6, pp. 713-722, 1998.

[5] Dia, H., & Rose, G., Development and Evaluation of Neural Network Freeway Incident Detection Models Using Field Data. *Transportation Research C*, Vol. 5, No. 5, 1997, pp. 313-331.

[6] Evanco, W.M., the Potential Impact of Rural Mayday Systems on Vehicular Crash Fatalities. Accident Analysis and Prevention, Vol. 31, 1999, pp. 455-462.

[7] Kim, K., Nitz, L., Richardson, J., & Li, L., Personal and Behavioural Predictors of Automobile Crash and Injury Severity. *Accident Analysis and Prevention*, Vol. 27, No. 4, 1995, pp. 469-481.

[8] Kweon, Y. J., & Kockelman, D. M., Overall Injury Risk to Different Drivers: Combining Exposure, Frequency, and Severity Models. Accident Analysis and Prevention, Vol. 35, 2003, pp. 441-450.

[9] Shankar, V., Mannering, F., & Barfield, W., Statistical Analysis of Accident Severity on Rural Freeways. *Accident Analysis and Prevention*, Vol. 28, No. 3, 1996, pp.391-401.

[10] Ossiander, E. M., & Cummings, P., Freeway speed limits and Traffic Fatalities in Washington State. Accident Analysis and Prevention, Vol. 34, 2002, pp. 13-18.

[11] Martin, P. G., Crandall, J. R., & Pilkey, W. D., Injury Trends of Passenger Car Drivers In the USA. Accident Analysis and Prevention, Vol. 32, 2000, pp. 541-557.

[12] Mayhew, D. R., Ferguson, S. A., Desmond, K. J., & Simpson, G. M., Trends In Fatal Crashes Involving Female Drivers, 1975-1998. Accident Analysis and Prevention, Vol. 35, 2003, pp. 407-415.

[13] Mussone, L., Ferrari, A., & Oneta, M., An analysis of urban collisions using an artificial intelligence model. *Accident Analysis and Prevention*, Vol. 31, 1999, pp. 705-718.

[14] Yang, W.T., Chen, H. C., & 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.

[15] Sohn, S. Y., & 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, February 2003, pp. 1-14.

[16] Tavris, D. R., Kuhn, E. M, & Layde, P. M., Age and Gender Patterns In Motor Vehicle Crash injuries: Importance of Type of Crash and Occupant Role. Accident Analysis and Prevention, Vol. 33, 2001, pp. 167-172.

[17] Kweon, Y. J., & Kockelman, D. M., Overall Injury Risk to Different Drivers: Combining Exposure, Frequency, and Severity Models. Accident Analysis and Prevention, Vol. 35, 2003, pp. 441-450.

[18] Zembowicz, R. & Zytkow, J. M., 1996. From Contingency Tables to Various Forms of Knowledge in Database. Advances in knowledge Discovery and Data Mining, editors, Fayyad, U. M. et al., AAAI Press/The MIT Press, pp.329-349.

[19] Akomolafe, O.P. (2004); predicting possibilities of Road Accidents occurring, using Neural Network. M. Sc. Thesis, Department of Computer Science, University of Ibadan.

Authors' Profiles

OLUTAYO V.A: Has M.Sc. in computer science from University of Ibadan, Nigeria. Post-graduate student for doctoral degree for computer science in University of Benin, Nigeria. Currently, an Assistant Lecturer in Computer Science Department Joseph Ayo Babalola University, Ikeji-Arakeji, Nigeria. Member, Computer Professionals Association of Nigeria (CPN). Major in Data Mining, and Service Oriented Computing.

ELUDIRE A.A: Has M.Sc., Ph.D. in Computer Engineering from Kiev Polytechnical Institute, Ukraine. Currently, a Senior Lecturer at the Computer Science Department, Joseph Ayo Babalola University, Ikeji-Arakeji, Nigeria. Member of IEEE, Nigeria Society of Engineer and Nigeria Computer Society (NCS). Major in Computer Networking and Computer Architecture and Organization.

How to cite this paper: Olutayo V.A, Eludire A.A,"Traffic Accident Analysis Using Decision Trees and Neural Networks", International Journal of Information Technology and Computer Science(IJITCS), vol.6, no.2, pp.22-28, 2014. DOI: 10.5815/ijitcs.2014.02.03

