Journal, Advances in Mathematical & Computational Sciences
Vol. 8 No. 3, September, 2020
www.mathematics-computationaljournal.info

Journal of Advances in Mathematical & Computational Sciences
An International Pan-African Multidisciplinary Journal of the SMART Research Group
International Centre for IT & Development (ICITD)
Southern University Baton Rouge, LA, USA
© Creative Research Publishers - Available online at
https://www.isteams.net/socialinformaticsjournal
DOI: dx.doi.org/10.22624/AIMS/MATHS/ V8N3P1
CrossREF Member Listing - https://www.crossref.org/06members/50go-live.html

Knowledge Discovery in Road Accidents Database: A Multidimensional Approach

Olutayo, V.A. & Eludire, A.A.
Department of Computer Science
Joseph Ayo Babalola University
Ikeji-Arakeji, Osun State, Nigeria
E-mails: vaolutayo@jabuledu.ng; aaeludire@gmail.com

ABSTRACT

Road accident statistics are collected and used by a large number of users and this can result in a huge volume of data which requires to be explored in order to ascertain the hidden knowledge. Potential knowledge may be hidden because of the accumulation of data, which limits the exploration task for the road safety expert and, hence, reduces the utilization of the database. In order to assist in solving these problems, this paper explores Decision Trees and multidimensional 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 from Ibadan to Lagos were collected from Nigeria Road Safety Corps. The data were pre-processed and organized into categorical form, which were also analyzed by Decision Trees algorithms. 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), Normalized Mean Square Error (NMSE), Accuracy Rate, True Positive, False Positive and Percentage correctly classified instances. Empirical results show that the regression Tress could classify accidents causes within reasonable accuracy. The multidimensional data analysis showed that the dark spot on the road is located between kilometers 10 – 20 where 60.2% of the accident took place. Tyre burst, loss of control and over speeding were there the major causes of accidents.

Keywords: Knowledge discovery, Decision Trees, accidents, data pre-processing, multidimensional data analysis.

Olutayo, V.A. & Eludire, A.A. (2020): Knowledge Discovery in Road Accidents Database: A Multidimensional Approach. Journal of Advances in Mathematical & Computational Sc. Vol.8, No. 3. Pp 13-22. DOI: dx.doi.org/10.22624/AIMS/MATHS/ V8N3P2. Available online at www.isteams.net/ mathematics-computationaljournal.

1. INTRODUCTION

Road traffic accidents are among the top leading causes of deaths and injuries of various levels. Nigeria is experiencing highest rate of such accidents resulting in death and various levels of injuries which call for serious concern. Highways in Nigeria, takes the lion's share of the risk having higher number of vehicles and traffic and the cost of these death, fatalities and injuries has a great impact on the socio-economic development of a Nigerian society. Truma is a major cause of morbidity and mortality and road accident is a special case of trauma that constitutes a major cause of disability and untimely death and loss of loved ones as well as family bread winners [1, 2, 3].

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 [4]. Statistics have also shown that mortality in road accidents is very high among young adults that constitute the major part of the work force [3,4, 5]. In actual fact, accidents kill faster than AIDS and it gives no preparatory time to its victims.

In order to combat this problem, various road safety strategies have been proposed and used. These methods mainly involve conscious planning, design and operations on roads. One important feature of this method is the identification and treatment of accident prone locations commonly called black spots [6]. 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.

The purpose of this research is to look at historical data of road accidents on one of the Nigeria's busiest roads from many dimensions 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. Shankar, et al. [3] 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. [4,6,5] used Poisson regression to analyzed 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. Martin et al. [5] analyzed the relationship be-tween road infrastructure and safety by using a cross-sectional time-series data base collected for all 50 U.S. states over 14 years.

The result suggested that as highway facilities are upgraded, there are reduced fatalities. Yang, et al. [6] used NN approach to detect safer driving patterns that have less chances of causing death and injury when a car crash occurs [5, 7]. 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. 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%. Osofisan et al. [8] employed Artificial Neural Network using Multilayer perceptron to predict likelihood of accident happening at particular location between the first 40 kilometers 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.

Abdelwahab et al. [7] 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%).

Olutayo et al. [1] did a work on traffic accident analysis using decision trees and neural networks, in their paper, a comparison of different Decision Tree algorithms and Artificial Neural Networks performance, were analysed on road accident data set. 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, in their work, 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.

2. METHODS

Knowledge discovery approach involves Data Mining (DM). Data Mining is an interactive process of discovering valid and novel, useful and understandable patterns or models in large databases. DM process is valid because it generalizes for the future; it is novel because it discovers what we do not know; useful because it enables us to take action; understanding because it gives us insight; iterative because the process takes multiple passes; and finally interactive because human beings are involved in the loop. The goals may be opaque and transparent. DM operations may be verification driven or discovery driven. This work is discovery driven.

We have chosen to employ DM tools in this research because they extend beyond traditional statistical approach and are well suited for high dimensional data as well as data with a mix of numeric and categorized data. DM models include Neural Network, Generic Algorithm, Decision Trees Algorithm, Probability networks and Expert systems technologies among others. The model adopted for this work is Decision Tree Algorithm because it has ability for both classification and regression problems. In addition, Decision Trees runs faster on categorical data. Data used in this work are categorical data, i.e., they are not numeric data. A decision tree is composed of hierarchically arranged nodes, growing from the top most nodes called root node to leaf node. So it can be thought as the tree growing upside down, splitting the data at each level to form new nodes.

The resulting tree comprises of many nodes connected by branches. Nodes that are at the end of branches are called leaf nodes and play a special role when the tree is used for prediction. That means each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this specific attribute. Decision Trees have emerged as a powerful technique for modelling general input / output relationships. They are tree – shaped structures that represents a series of roles that lead to sets of decisions.

They generate rules for the classification of a dataset and a logical model represented as a binary (two – way split) tree that shows how the value of a target variable can be predicted by using the values of a set predictor variables. The heart of any DM 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 on the first 40 kilometres from Ibadan to Lagos. It consists of label-variables as given in table 1. 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

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
5		Location	LOC1 LOC2 LOC3	Categorical

3. PERFORMANCE MEASUREMENT

The performance measures are used to measure the performance of the trees network for a particular set. Those used include Mean Absolute Error (MAE), Accuracy Rate, True Positive, False Positive and Percentage correctly classified instances. The MSE is used to determine how well the network output fits the desired output. MSE is defined as:

$$MSE = \frac{\sum_{j=0}^{p} \sum \mathrm{N}(dy-yij)2}{NP} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where P = number of output processing elements
N = number of exemplar in the data set network output for exemplar i at processing element
j = desired output for exemplar I at processing element j.

The correlation coefficient, (r)

given as:

$$r = \frac{\frac{\sum i(x_i - \dot{x})}{N}}{\frac{\sqrt{\sum(d_i - d)^2}}{N}\frac{\sqrt{\sum(x_i - \dot{x})}}{N}}$$

The Percent Error is defined as:

$$\%\ Error = \frac{100}{\mathrm{NP}} \sum_{j=0}^{P} \sum_{I=0}^{N} \frac{dy_{ij} - dd_{ij}}{dd_{ij}}$$

where *P* = number of output processing elements
N = number of exemplar in the data set
dyij = de-normalized network output for exemplar I at processing element j
ddij = de-normalized desired output for exemplar I at processing element.

4. Multidimensional analysis of data

The multi-dimensional analysis results are presented in Tables 3 to 7. Table 3 shows that the dark spot on the first 40 kilometers of Ibadan-Lagos express way is located between kilometers 10 to 20, that is, Location 2 where 60.2% of the accidents took place. All the accidents caused by wrong overtaking took place in location 2. No accident caused by robbery took place within the first 9 kilometers of the road. About one third of the accidents that occurred in location 1 are caused by loss of control, followed by over speeding and dangerous/careless driving. Thus majority of the accidents that occurred in Location 1 can be attributed to the drivers. A close look at table 3 shows that even though tyre bursts account for 38.3% of all accidents only 0.8% of this occurred in location 1. Therefore, it may be possible to reduce road accidents in location 1 if drivers have the right attitude of driving on the highway.

Apart from obstruction and broken shaft, location 2 has a fair share of all causes of accidents with tyre burst contributing the highest, followed by over speeding and loss of control respectively. 71.3% of tyre burst, 50% of loss of control and 58.1% of over speeding occurred in location 2. About 61.9% of the other factors in the cause of accidents also occurred in location 2. Going by the fact that only 2.1% of tyre burst occurred in location 1, the issue of tyre burst in location 2 may be as a result of other extraneous causes. It may therefore be necessary to further investigate location 2 with respect to reducing accidents in this spot. The major cause of accidents in location 3 is also tyre burst. However, unlike location 2, loss of control and over speeding are considerably reduced. No accident was caused by wrong overtaking, dangerous/careless driving, pushing by another car, broken spring, and brake failure in this location. Location 3 recorded the least number of accidents in spite of the fact that all the accidents caused by tree obstruction and broken shaft took place within it.

Most of the accidents occurred during the day that is 41.4% in the morning and 42.1% in the afternoon with only 1.6% occurring at night. Accidents due to robbery attack were evenly distributed among the three periods of morning, afternoon and evening. No accident due to robbery occurred at night. Night accidents were caused only by loss of control and tyre burst. There were more accidents during the dry season than the wet season. No accident occurred as a result of tree obstruction, pushing by another car, broken shaft or broken spring during the wet season.

Big vehicles were more involved in accidents than small vehicles in the ratio 59.9:40.1. These were in accidents caused by tyre burst and loss of control, while small vehicles were more prone to dangerous and careless driving as well as over speeding

4. CONCLUSION

This work employed Decision Trees algorithm as well as multi-dimensional analysis to analyze the data of accidents, which occur at a particular location, namely between the first 40 kilometers along the Ibadan-Lagos Express road. The historical data used were collected from the records of the FRSC and the Police along this highway. The study revealed that given enough historical categorical data, Decision Trees algorithms can predict black spots on express ways with astonishing accuracy. Location two has 77.8 percent correctness with a low mean square error for this reason.

Sensitivity analysis was also carried out to extract redundant input and finally, the causes of accidents were examined to discover the significant input to accident occurrence. Multidimensional analysis also revealed that the dark spot of the road is located within the first 10 - 20 kilometer stretch of the road. The condition of the road around this region should be critically checked and road safety activities should be intensified in order to reduce carnage on the Ibadan-Lagos express road.

Tyre burst, loss of control and over speeding were the three major causes of accidents. Tyre burst which represents the major cause of accidents may be as a result of extraneous factors that may require further investigation. The data input of location was ranged; further research can be carried out with a continuous value of locations. Also one of the input variables that can affect the occurrence of accident at a particular location is the topography of the location which

Table 2: Performance measure of the best Model

Performance	LOCATION(3)	LOCATION(2)	LOCATION(I)
MSE	0.188687417	0.22273544	0.160923237
NMSE	1.091691458	0.890941713	0.802140435
MAE	0.36017332	0.43966329	0.291889134
Min Abs Error	0.017737681	0.032159686	0.016360449
Max Abs Error	0.963505924	0.926017717	0.848296911
R	0.34129631	0.41751817	0.553423268
Percent Correct	50	77.77777863	40

Table 3: Causes of accidents from location dimension

Location	Causes of Accident by %													Total
	A	B	C	D	E	F	G	H	I	J	K	L	M	
LI	0	3.1	7.0	0.8	4.7	0.0	0.0	0.0	0.8	2.3	0.8	1.6	0.0	21.1
L2	3.1	3.9	7.8	27.3	8.6	0.0	0.8	0.0	0.8	0.8	1.6	3.9	1.6	60.2
L3	0.0	0.0	0.8	10.2	1.5	0.8	0.0	2.3	0.0	0.0	1.5	0.8	0.8	18.7
Total	3.1	7.0	15.6	38.3	14.8	0.8	0.8	2.3	1.6	3.1	3.9	6.3	2.4	100.0

Table 4: Causes of accident from time dimension

Time	Causes of Accident by %													Total
	A	B	C	D	E	F	G	11	1	J	K	L	M	
Morning	0.8	3.8	4.7	12.5	5.5	0.0	0.8	1.5	1.6	0.8	3.1	5.5	0.8	41.4
Afternoon	0.8	1.6	9.3	18.7	7.0	0.0	0.0	0.8	0.0	2.3	0.8	0.0	0.8	42.1
Evening	1.5	1.6	0.8	6.3	2.3	0.8	0.0	0.0	0.0	0.0	0.0	0.8	0.8	14.9
Night	0.0	0.0	0.8	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.6
Total	3.1	7.0	15.6	38.3	14.8	0.8	0.8	2.3	1.6	3.1	3.9	6.3	2.4	100.0

Table 5: Causes of accidents from season dimension

Season	Causes of Accident by %													Total
	A	B	c	D	E	F	G	H	1	J	K	L	M	
Wet	0.8	2.3	6.3	21.1	3.9	0.0	0.0	0.0	0.0	0.8	1.6	3.9	1.6	42.3
Dry	2.3	4.7	9.3	17.2	10.9	0.8	0.8	2.3	1.6	2.3	2.3	2.4	0.8	57.7
Total	3.1	7.0	15.6	38.3	14.8	0.8	0.8	2.3	1.6	3.1	3.9	6.3	2.4	100.0

Table 6: Causes of accidents from type dimension

Type of Vehicle	Causes of Accident by %													Total
	A	B	c	D	E	F	*G*	H	*I*	J	*K*	L	M	
Small	1.6	4.7	4.7	9.4	9.4	0.8	0.8	0.0	0.0	1.6	3.1	3.2	0.8	40.1
Big	1.5	2.3	10.9	28.9	5.4	0.0	0.0	2.3	1.6	1.5	0.8	3.1	1.6	59.9
Total	3.1	7.0	15.6	38.3	14.8	0.8	0.8	2.3	1.6	3.1	3.9	6.3	2.4	100.0

Table 7: Percentage summary of causes of accidents

Code	Causes of Accident	% Contribution
A	Wrong Overtaking	3.1
B	Dangerous/careless Driving	7.0
C	Loss of Control	15.6
D	Tvre Burst	38.3
E	Over Speeding	14.8
F	Obstruction	0.8
G	Pushed by another Vehicle	0.8
H	Broken Shaft	2.3
I	Broken Spring	1.6
J	Brake Failure	3.1
K	Road Problem	3.9
L	Unknown	6.3
M	Robbery Attack	2.4
	TOTAL	100.00

REFERENCES

[1] Olutayo, V.A. & Eludire, A. A. (2014) Traffic Accident Analysis Using Decision Trees and Neural Networks. *I.J. Information Technology and Computer Science,* Volume 02, pp. 22-28 Published Online in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijitcs.2014.02.03

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

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] Osofisan, A. O, Akomolafe, O. P., & Akinola, S. O. (2005) Discovering Knowledge in Road Accident Database. International Center for Mathematical & Computer Sciences. Lagos, Nigeria. Volume 2, pp.31-43

[9] Balogun J.A. Abereoje O.K.(1987) Pattern of road traffic accident cases in a Nigeria University Teaching Hospital between 1987 and 1990, J,Trop Med Hyg 199; 95:29.