Predicting Causes of Traffic Road Accidents Using Multi-class Support Vector Machines

Elfadil A. Mohamed

Department of Management Information Systems, College of Business Administration, Al Ain University of Science and Technology, Al Ain, United Arab Emirates

Abstract *- Road traffic accidents have caused a myriad of problems for many countries, ranging from untimely loss of loved ones to disability and disruption of work. In many cases, when a road traffic accident occurs that results in the death of both drivers of the vehicles involved in the accident, there are some difficulties in identifying the cause of the accident and the driver who committed the accident. There is a need for methods to identify the cause of road traffic accidents in the absence of eyewitnesses or when there is a dispute between those who are involved in the accident. This paper attempts to predict the causes of road accidents based on real data collected from the police department in Dubai, United Arab Emirates. Data mining techniques were used to predict the causes of road accidents. Results obtained have shown that the model can predict the cause of road accidents with accuracy greater than 75%.*

Keywords—component; road traffic accident, data mining, multi-class SVMs

1 Introduction

Road traffic accidents (RTAs) are currently ranked ninth globally among the leading causes of disability-adjusted life years lost and the ranking is projected to rise to third by 2020 [1]. A study conducted by Bener et al. [2] indicates that road traffic fatalities are second only to cardiovascular disease in the list of major causes of death. About 90% of the disability-adjusted life years lost worldwide due to road traffic injuries occur in developing countries. In recent years, high rates of serious RTAs have been reported in several Arabian Gulf countries, including the United Arab Emirates. UAE is a young, wealthy country with a number of vehicles on the road that is continuously increasing. The rate of RTAs is relatively high in the UAE and generally causes more serious trauma than other accidents, which is reflected in a high number of fatal and serious injuries.

RTA prediction is an essential problem in traffic safety control. It is acknowledged that the success of traffic safety and highway improvement programs hinges on the analysis of accurate and reliable traffic accident data. For the successful completion of traffic safety controls, robust computational methods for predicting RTA are seriously needed; therefore, the subject was intensively studied by researchers around the globe. RTAs that lead to the death of all on board the vehicles involved in the accident leave the traffic police without eyewitnesses to question in order to determine the cause of the accident. Even when an accident does not result in death, there might be disputes between those who are involved in the accident to know who is the victim and who is the offender. The police department might experience some difficulties in identifying the real offender. Methods are needed to predict the cause of the accident and the offender in RTAs. On the other hand, preventing accidents from happening is a major challenge. Computation and data mining methods are greatly needed to predict the possible causes of RTAs.

The main objective of this study is to design effective data mining methods to investigate and predict the cause of RTAs in one of the Gulf countries; namely, United Arab Emirates (UAE). Real data were obtained from the Dubai police department and were used for building a multi-class support vector machine for predicting the possible cause of TRAs. The remainder of the paper is organized as follows: Section II discusses the literature related to the prediction of number of RTAs and the forecasting of the cause thereof. Section III explains the methodology used. Section IV discusses the experimental work and results. Finally, Section V provides a conclusion of the work.

2 Related Works

Most of the literature related to RTAs is centered on the prediction of the number of road accidents; for example, Huilin and Yucai [3] used neural networks for the prediction of traffic accidents. Yisheng et al. [4] have investigated the use of the k-nearest neighbor method for identifying the more likely traffic conditions leading to traffic accidents, while considering the joint effects of accident precursors on traffic accident occurrences and controlling the geometry and environmental factors. For forecasting of RTA, Qing-wei et al. [6] used a method that combine support vector regression (SVR) and particle swarm optimization (PSO). The experimental results indicated that the proposed PSO-SVR method has better performance accuracy than back propagation neural network in traffic accident forecasting. Mathematical models have been used for the estimation of the number of RTAs. A novel composite grey back-propagation neural network model was proposed by Zhu [7] for the estimation of the number of RTAs. The proposed model

showed an improvement in the forecasting accuracy of the number of RTAs.

With the availability of data in an electronic form, data mining techniques have widely been used in road traffic accident analyses. Classification and clustering techniques were used in [8] for the prediction of traffic incidents. Spatial data mining for the analysis of traffic accidents is introduced in [9]. Recent study that addresses issues related to the use of data mining techniques for predicting the likelihood occurrence of road traffic accident on highways, the likely cause of the accident, and accident-prone locations can be found in [5].

The development of data mining models that predict the number and cause of RTAs has been studied extensively (see, for example, [3, 6, 7] for predicting the number of RTAs and [5] for predicting the cause of RTAs). However, most of the above-mentioned data-mining models suffer from low prediction accuracy of the prediction and, in most of the cases, it is due to a poor pre-processing step. New data mining methods are needed with the aim of improving the accuracy of predictions through the understanding of the fuzziness of the datasets, solid pre-processing (handling of missing entries, unbalance data issues, wrong entries, evaluating attributes related to RTAs, etc.), and usage of powerful machine learning techniques such as support vector machines (SVMs).

Research work studying and predicting the cause of RTAs in the Gulf region is quite limited. In Saudi Arabia, for instance, Ageli and Zaidan [10] used an autoregressive distributed lag ADRL model for the analysis of RTAs. Bener et al. [1] explored the pattern of RTAs and their causes in the state of Qatar. Bener and Crundal [2] have investigated data concerning RTAs and road user behavior in UAE. However, no published research work focuses on studying and addressing issues related to the prediction of either the number of RTAs or the cause of RTAs in the UAE.

Data pre-processing is an important step in data mining. The main purpose of data preprocessing is to improve the quality of the data, which leads to the improvement of the mining results. According to Han et al. [11], there are several data pre-processing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in data. Data integration merges data from multiple sources into a coherent data store such as a data warehouse. Data reduction can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering. Finally, data transformation may be applied, where data are scaled to fall within a smaller range. Researchers have long recognized the importance of data pre-processing; for example, Kotsintis et al. [12]'s work address issues of data pre-processing that can have significant impact on the generalization performance of machine learning algorithms. For improving the efficiency of data preprocessing, Chen and Liu [14] proposed an improved data cleaning method.

Missing values can be handled by either ignoring the record, using global constants to fill the missing value, using measures of central tendency for the attributes, or by using the most probable value to fill in the missing values [11]. Kumar and Kalia [13] used the average value to fill the missing values. For the handling of missing values, Higashijima et al. [15] proposed the use of a regression tree imputation method. The proposed method achieved high accuracy compared to the not-pre-processed and linear interpolation method. In this paper, several valid pre-processing methods are used to improve the quality of RTA data before the development of the multi-class SVM step.

3 Method

In this case, our solution will follow the typical data-mining framework, which consists of three main steps: preparing the data (pre-processing), mining patterns, and post-processing. These steps will be described in the following sections. Figure 1 illustrates the main components of the method we used in this research.

3.1 Preparing and preprocessing the accident data

3.1.1 Data collection:

The data set used in this study is collected and retrieved from the Dubai Police Department, UAE. The traffic police department gathers and records the accident data using a traffic accident information system. The data consists of 7,048 entries and seven different attributes related to the drivers (age, nationality, and license) and the vehicles (type, year of make, etc.) involved in the accidents besides the location where the accidents took place.

The seven attributes used are locations (914 different locations), DEGINJ (4 different types), gender (M/F), age (67 different ages), country the driver belongs to (31 different countries), vehicle type (9 different types), and year the vehicle is made (33 different years of made).

3.1.2 Data preprocessing:

The issue of data quality has a direct relevance on the quality of the data mining results. Although almost everyone accepts the importance of data quality (please refer to section II), in reality, it is not always rigorously controlled. Traffic data is no exception and it suffers from unknown or missing entries, consistency, completeness, redundancy, etc. In this particular case, our data has missing entries with respect to the car's manufacturing year and the driver's gender and should be handled before the development of the data mining model. To handle the missing values, we adopted the "ReplaceMissingValues" method available in Weka machine learning software, which uses modes and means to identify the missing entries.

To ensure that all our attributes are meaningful, a feature selection is used to assess the relevance of each attribute. In this case, we focus on using a feature selection method that is based on filtering. Filtering algorithms use independent search and evaluation methods to determine the relevance of features variables to the data-mining task. In this case, we used the

"GainRatioAttributeEval" method available in Weka to evaluate the worth of an attribute by measuring the gain ratio with respect to the class.

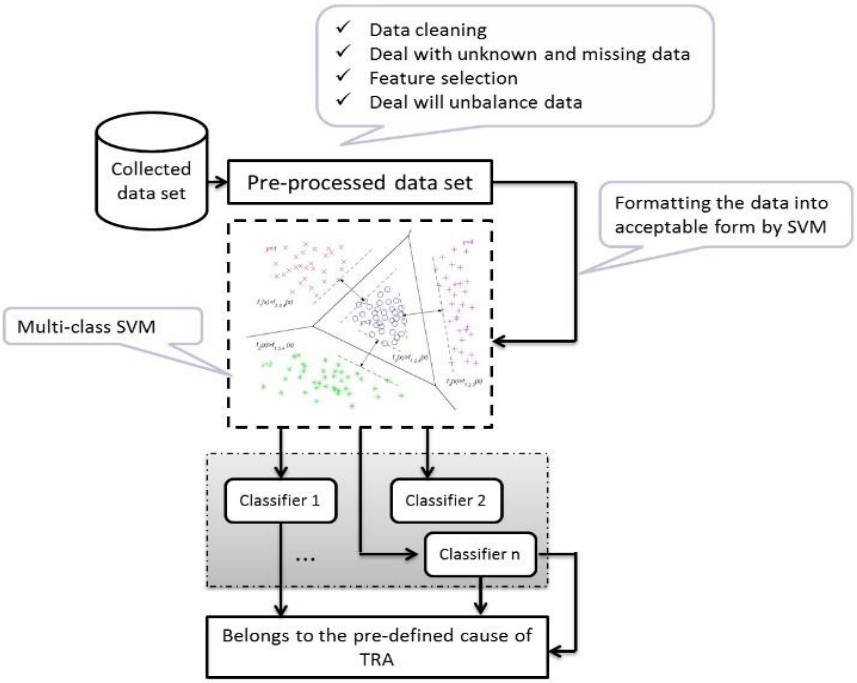

Fig. 1. The framework of predicting the cause of road traffic accident using multi-class SVMs.

3.1.3 Mining patterns

Once the data is pre-processed, a sensible data-mining task must be designed to comply with the objectives of predicting the 21 causes of road accidents. This problem can be handled by utilizing a multi-classification technique; therefore, Support Vector Machines (SVM) [20, 21] was selected. The idea of the SVM algorithm is to map the given training set into a possibly high-dimensional feature space and attempting to locate in that space a hyperplane that maximizes the distance separating the positive from the negative examples [16, 17].

The SVM algorithm addresses the general problem of learning to discriminate between positive and negative examples of a given class of n-dimensional vectors [18, 22]. In order to discriminate among the 21 causes of road accidents, the SVM learns a classification function from a set of positive

Examples $\mu+$ and set of negative examples $\mu-$. The classification function takes the form:

$$f(x) = \sum_{i:x_i \in \mu+} \lambda_i K(x, x_i) - \sum_{i:x_i \in \mu-} \lambda_i K(x, x_i) \qquad (1)$$

where the non-negative weights λ_i are computed during training by maximizing a quadratic objective function and the function $K(x, x_i)$ is called a kernel function. Any accident case x is then predicted to be positive if the function $f(x)$ is positive. More details about how the weights λ_i are computed and the theory of SVM can be found in [16, 17].

3.1.4 Post-processing patterns

Following the classification step, it is important to evaluate the patterns detected by the SVM. Several evaluation measures are used in this study, such as Precision (Pr = $\frac{TP}{TP+FP}$), Recall (Re = $\frac{TP}{TP+FN}$), F1 measure ($2*\frac{\Pr \cdot \mathrm{Re}}{\Pr + \mathrm{Re}}$) and Accuracy (Ac = $\frac{TP+TN}{All}$, where TP, TN, FP, FN, and All are defined as:

- TP: related cause of road accident classified as "related."
- TN: unrelated cause of road accident classified as "unrelated."
- FP: related cause of road accident classified as "unrelated."
- FN: unrelated cause of road accident classified as "related."
- All: total number of causes of road accidents.

4 Experimental Work and Results

The experimental work began with the exploration and the preparation of the dataset. Several missing entries were found, particularly under the car's manufacturing year (2.23%) and gender (18.25%) attributes. The ReplaceMissingValues method was applied without referring to a particular class. Once the missing data entries were handled, the seven attributes were analyzed and the GainRatioAttributeEval method reveals that the location, vehicle type, and the driver's country have a strong relationship with the cause of the accident. Similarly, there is no evidence suggesting that gender has a link to the cause of accidents. Details of the attribute evaluation are summarized in Table 1.

TABLE 1

ATTRIBUTE EVALUATION

Attribute	Rank	Gain ratio
Location	1	0.1932
Vehicle type	2	0.0704
Country the driver belongs to	3	0.0494
Age	4	0.0353
DEGINJ (level of injury)	5	0.0317
Year the vehicle was made	6	0.0172
Gender	7	0

One other observation inferred from the data exploration is the fact that the data is unbalanced. The distribution of causes of accidents is shown in Figure 2. It is quite obvious to see that most of the accidents took place in the UAE due to an absence of attention/consideration for other drivers or excessively speeding.

From a data-mining point of view, this data requires balancing; therefore, a resampling method with a random seed equal to 1 was used. The resampling in this case produces a random subsample of a dataset using replacements.

Once the preprocessing step is completed and the dataset is prepared, a multiclass SVM was used. The Library for Support Vector Machines (LibSVM) implemented by Chih-Chung Chang and Chih-Jen Lin [23] was used. One of the significant parameters needed to tune the SVM is the choice of the kernel function. The kernel function allows the SVM to locate the hyperplane in a highly dimensional space that effectively separates the training data [16, 17]. The Gaussian Radial Basis function is used as it allows pockets of data to be classified, which is more powerful than just using a linear dot product [16].

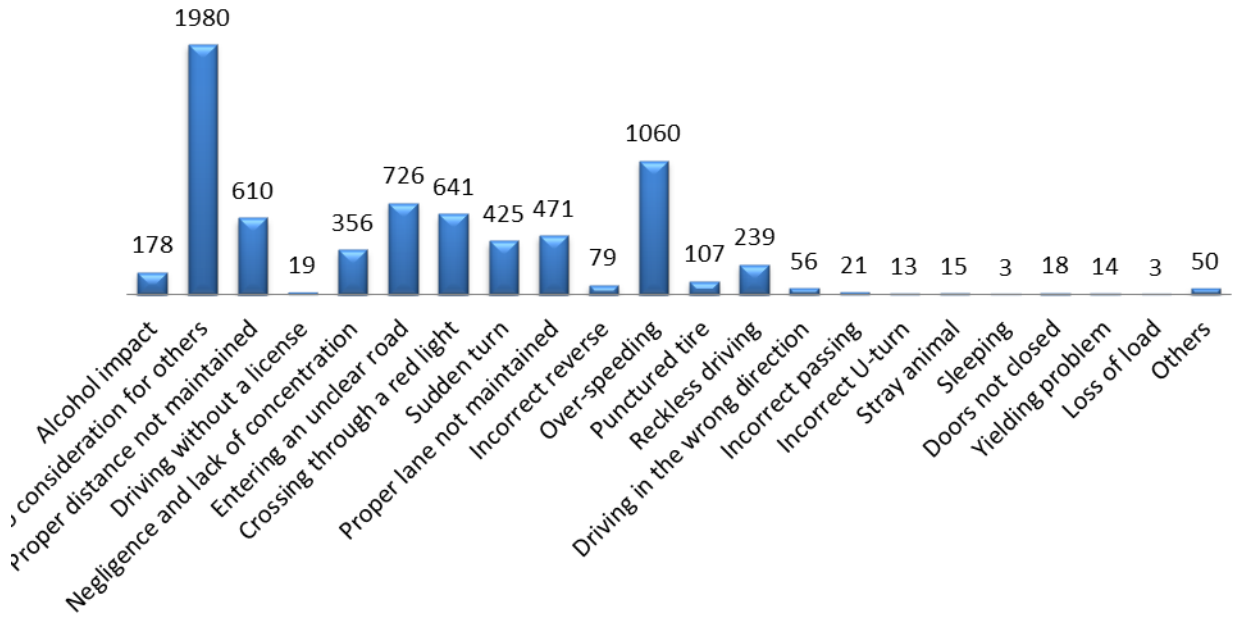

Fig. 2. Distribution of the causes of accidents in the UAE

Using 10-fold cross-validation, the detailed classification results are shown in Table 2. The overall Precision, Recall, F1 measure, and Accuracy are 0.767, 0.754, 0.752 and 75.395, respectively. Apart from the "Stray animal," "Yielding problem" and "Fall of the load" causes, the remaining 18 causes were detected with high accuracy. The three mentioned causes were had a very limited number of samples—15, 14, and 3, respectively.

TABLE 2

CLASSIFICATION DETAILS OF CAUSE OF ACCIDENTS

Cause of accident	Precision	Recall	F1-Measure
Alcohol impact	0.807	0.669	0.731
No consideration for others	0.693	0.896	0.782
Proper distance not maintained	0.718	0.636	0.674
Driving without a license	0.848	0.824	0.836
Negligence and lack of concentration	0.8	0.682	0.736
Entering an unclear road	0.858	0.749	0.799
Crossing through a red light	0.899	0.762	0.825
Sudden turn	0.84	0.642	0.728
Proper lane not maintained	0.781	0.647	0.708
Incorrect reverse	0.839	0.635	0.723
Over-speeding	0.687	0.77	0.726
Punctured tire	0.724	0.598	0.655
Reckless driving	0.901	0.628	0.74
Driving in the wrong direction	0.977	0.764	0.857
Incorrect passing	0.923	0.75	0.828
Incorrect U-turn	1	0.81	0.895
Stray animal	0	0	0
Sleeping	0.8	1	0.889
Doors not closed	0.5	0.5	0.5
Yielding problem	0.5	0.3	0.375
Loss of load	0	0	0
Others	0.722	0.433	0.542
Average	**0.767**	**0.754**	**0.752**

5 Conclusions

In this paper, a multi-class support vector machine model was developed to enable the prediction of the cause of RTAs in UAE. Several preprocessing methods are used to improve the quality of the traffic data. The accuracy achieved by the developed model is approximately 75%, which is quite acceptable despite the fact that further accuracy improvements are required. The developed model can be used by the UAE traffic police department as a tool for predicting the future causes of RTAs, as well as the offending driver, in the case of the absence of eyewitnesses or when there is a dispute between those who are involved in the accident. The model can assist in avoiding accidents. The good accuracy achieved in this study suggested that there are strong combinations (patterns) of attributes that could lead to possible common causes of accidents. Taking these combinations in consideration, the traffic authorities could communicate and warn drivers to be more alert. Moreover, the analysis of the RTA data has revealed that the dominant causes of RTA in the UAE are typically due to neglecting other vehicles on the road or over-speeding. These causes of RTA are more related to the drivers' behavior. Traffic police authorities could conduct campaigns and awareness sessions to educate drivers in how to avoid these two causes of RTAs. Other methods of controlling over-speeding should be implemented. In the current study, only seven features/attributes were used. In the future, more valuable features describing the cause of RTA should be collected and analyzed.

6 Acknowledgment

The author would like to thank Mr. Omar Alnakbi, an employee at Roods & Transport Authority, Dubai, for his effort in providing the data which were used for testing the model.

7 References

[1] BENER, "The neglected epidemic: Road traffic accidents in a developing country, State of Qatar," International Journal of Injury Control and Safety Promotion, Vol. 12, No. 1, March 2005, pp. 45 – 47.

[2] Bener and D. Crundall, "Road traffic accidents in the United Arab Emirates compared to Western countries," Advances in Transportation Studies in an international Journal Section A 6, 2005.

[3] F. Huilin and Z. Yucai, "The Traffic Accident Prediction Based on Neural Network," Second International Conference on Digital Manufacturing & Automation, 2011.

[4] L. Yisheng, T. Shuming , and Z. Hongxia, "Real-time Highway Traffic Accident Prediction Based on the k-Nearest Neighbor Method," International Conference on Measuring Technology and Mechatronics Automation, 2009.

[5] T. Akomolafe and A. 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.

[6] Z. Qing-wei, F. Ai-Ying, and X. Zhi-Hai, "Application of Support Vector Regression and Particle Swarm Optimization inTraffic Accident Forecasting," International Conference on Information Management, Innovation Management and Industrial Engineering, 2009.

[7] X. Zhu, "Application of Composite Grey BP Neural Network Forecasting Model to Motor Vehicle Fatality Risk," Second International Conference on Computer Modeling and Simulation, 2010.

[8] Pan, U. Demiryurek, C. Shahabi, and C. Gupta, "Forecasting Spatiotemporal Impact of Traffic Incidents on Road Networks," IEEE 13th International Conference on Data Mining, 2013.

[9] W. Jinlin, C. Xi, Z. Kefa, W. Wei, and Z. Dan, "Application of Spatial Data Mining in Accident Analysis System," International Workshop on Education Technology and Training & 2008 International Workshop on Geoscience and Remote Sensing, 2008.

[10] M. Ageli and A. Zaidan, "Road Traffic Accidents in Saudi Arabia: An ADRL Approach and Multivariate Granger Causality," International Journal of Economics and Finance, Vol. 5, No. 7, 2013.

[11] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques. Morgan Kaufmann: Waltham, MA 02451, USA, 2012.

[12] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, " Data Preprocessing for Supervised Leaning," International Journal of Computer Science, Volume 1, No. 2, 2006.

[13] M. Kumar and A. Kalia, "Preprocessing and Symbolic Representation of Stock Data," Second International Conference on Advanced Computing & Communication Technologies, 2012.

[14] X. Cheng and H. Liu, "Research on Data Preprocessing Technology in Safety Equipment Linkage System," International Conference on Computational and Information Sciences, 2013.

[15] Y. Higashijima, A. Yamamoto, T. Nakamura, M. Nakamura, and M. Matsuo, "Missing Data Imputation using Regression Tree Model for Sparse Data collected viaWide Area Ubiquitous Network," 10th Annual International Symposium on Applications and the Internet, 2010.

[16] N. M . Zaki, S. Deris, and R. Illias, "Feature Extraction for Protein Homology Detection using Hidden Markov Model combining Scores," International J. of Computational Intelligence and Applications. Vol. 4, pp: 1-12, 2004.

[17] A. Abdalla, S. Deris, and N. M. Zaki, "Breast Cancer Detection Based on Statistical Textural Features Classification," International Conference on Innovations in Information Technology, 728-730, UAE, 2007.

[18] N. M. Zaki, and W. El-Hajj, "Predicting Membrane Proteins Type Using Inter-domain Linker Knowledge," The 2010 International Conference on Bioinformatics and Computational Biology (BIOCOMP'10), 12-15 July 2010, Las Vegas, US.

[19] N. M. Zaki, S. Deris, and H. Alashwal, "Protein-Protein Interaction Detection Based on Substring Sensitivity Measure," International J. of Biomedical Sciences, Vol. 1, pp: 148-154, 2006.

[20] V. N. Vapnik, Statistical Learning Theory. Wiley, 1998.

[21] N. Cristianini, and J. Shawe-Taylor, An introduction to Support Vector Machines. Cambridge, UK: CambridgeUniversity Press, 2000.

[22] N. M. Zaki, S. Deris, and R. M. Illias, "Simple Representation of Protein Sequence for detecting homology," International Conference on Artificial Intelligence. Las Vegas, USA, 27-30 June 2005.

[23] R. E. Fan, P. H. Chen, and C. J. Lin, "Working set selection using second order information for training SVM," Journal of Machine Learning Research 6, 1889-1918, 2005.