Evaluation of the Performance of Random Forests Technique in Predicting the Severity of Road Traffic Accidents

Salah Taamneh[1,*], Madhar Taamneh[2]
[1]The Hashemite University, Computer Science Department, Zarqa 13115, Jordan
Taamneh@hu.edu.jo
[2]Yarmouk University, Civil Engineering Department, Irbid, Jordan
mtaamneh@yu.edu.jo

Abstract. Traffic accidents in the Middle East are a primary concern for governments and local communities owing to the large numbers of fatalities, injuries and economic losses. Many analytical methods have been used in the literature to analyze the accidents database. One of the recent methods in this domain is the data-mining techniques. In this paper, we evaluate the performance of a well-known data mining technique called Random Forests (RF) in predicting the severity of road accidents based on 5973 accidents occurred in Abu Dhabi over a period of 6 years (2008-2013). The factors studied in this paper include: five accident-related attributes (year, day, time, reason of accident, and accident type), six driver-related attributes (gender, nationality, age, seat belt use, casualty status, degree of injury), and five road-related attributes (lighting, road surface, speed limit, lane numbers, and weather). The severity of the accident was classified into one of four classes (Minor, Moderate, Severe, and Death). RF was then used to build a prediction model using 10-fold cross validation method. The overall model predication performance was 68.5%. The generated model was found to perform poorly on the underrepresented classes (Death and Severe). As a result, the original data was transformed into a balanced data set using Minority Oversampling Technique (SMOTE). The performance of RF on the balanced data was 78.19% with 14% improvement. In order to validate the performance of the RF model, an ordered probit model was also used as a comparative benchmark. The accuracy of the ordered probit model was 59.5%, and 34% for the original and balanced data sets respectively. It was obvious that RF technique outperforms the ordered probit method in predicting the severity or road traffic accidents.

1 Introduction

Traffic accidents in the Middle East are a primary concern for governments and local communities due to the large numbers of fatalities, injuries and economic losses they cause. The analysis of road accidents can highlight the most important factors that play a role in the occurrence of road accidents, thus helping decision makers to take appropriate measures to prevent the occurrence of such accidents in the future. Many analytical methods have been used in the literature to analyze the accidents database. In recent years, there has been a growing interest in using data mining techniques to uncover hidden patterns in large multidimensional datasets. Such techniques can determine the interactions between variables that would be impossible to establish directly, using ordinary statistical modeling techniques [1] [2]. Specifically, data mining techniques were used effectively to identify the most influential factors affecting the severity of road traffic accidents [3] [4] [5] [6] [7]. In this paper a well know data mining technique called Random Forest (RF), was used to predict the injury severity of traffic accidents in Abu Dhabi, UAE.

RF is a data mining technique that was introduced by [8]. It is mainly used for classification and regression. Classification methods aim to identify to which category a new observation belongs based on a training set of data containing instances whose category membership is known. Decision tree algorithms are used to build classification models that predict the value of a target attribute based on the input attributes. The standard decision trees construct classification models in the form of trees. Each interior node in these trees represents one of the input variables, and it has a number of branches equal to the number of possible values of that input variable. Each leaf node holds a value of the target attribute. The leaf node represents the decision made based on the values of the input variables from the root to the leaf. The main issue with the standard decision trees is that they tend to overfit their training set [9]. In other words, the generated classifiers perform very well on the training set but poorly on new observations. RF technique was designed to solve this problem by averaging multiple deep decision trees trained on different parts of the same training set [10]. Random Forests apply the bootstrapping aggregating technique to reduce variance and helps avoid overfitting [11].

The RF technique has received increasing attention over the last few years due to the accurate classification results and the speed of processing. It has been applied in a variety of fields ranging from gene classification, land cover classification, disease risk prediction, image classification, and weather forecasting. Several studies were conducted to evaluate the performance of RF in classifying the severity of road accidents. One study evaluated the performance Naive Bayes Bayesian classifier, AdaBoostM1 Meta classifier, PART Rule classifier, J48 Decision Tree classifier and Random Forest Tree classifier for classifying the type of injury severity of various traffic accidents [12]. The results showed the RF outperformed the other methods. RF technique was also employed by [13] to identify traffic/highway design/driver vehicle information significantly related with fatal/severe crashes on urban arterials for different crash types. It was found that RF can help identify the roadway locations were severe crashes tend to occur. A mixed logic approach to investigate the effect of driver's age and side of impact on crash severity along urban freeways was proposed by [14]. Rather than starting with all possible variables, and then reducing one-at-a-time, this approach uses RF to assemble a screened list of variables to be entered in the mixed logit model. Fifty trees was used to grow the forest, which was sufficient to yield reliable results. RF was also used by [15] to rank the importance of the drivers/vehicles/environments characteristics on crash avoidance maneuvers. [16] Used RF to analyze the importance of several variables on traffic crash severity. The results revealed that the speed limit and single/multi-vehicle collision have the largest influence on the severity of traffic crashes. Based on this result, a series of Bayesian ordered logistic model for low/medium/high speed roads and single/multiple collision cases was built.

2 Methodology

In this paper, the performance of the Random Forest technique in predicting the severity of road traffic accidents was evaluated. The output of this technique is a decision tree that is used to predict the value of a target attribute (i.e., the severity of an accident) based on the input attributes. Traffic accident data used to build the prediction model were obtained from Abu Dhabi Emirate for a six-year period (2008-2013). In order to validate the performance of RF technique, an ordered probit model was also used as a comparative benchmark.

2.1 Data collection and preprocessing

Traffic accident data were obtained from Abu Dhabi Emirate for a 6-year period (2008–2013). The total number of accident records obtained for this period was 5973. For each accident record, 48 different attributes were collected at the time of the accident. To investigate the influence of factors on the severity of crashes, the dependent variable (degree of injury) was categorized into four levels, of death, severe, moderate and minor accidents. Descriptions of each category are presented in Table 1. The proportions of each category are shown in Figure 1. Of 5973 traffic accidents, 59% were involved in minor accidents, 31% were involved in moderate accidents, 7% were involved in severe accidents and 3% were involved in fatal accidents.

Table 1. Class labels

No.	Accidents Categories	Categories Description
1	Death	One or more person dies within 30 days of the accident.
2	Sever	A person is injured and requires intensive care.
3	Moderate	One or more persons injured and detained in hospital for more than twelve hours.
4	Minor	All persons involved either not detained in hospitals or detained for not more than twelve hours.

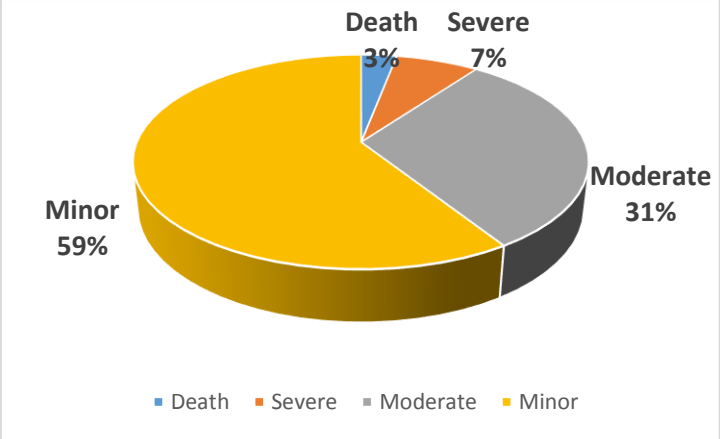

Fig. 1. Traffic Accident Severity Proportion

Traffic accident data were obtained in the form of an Excel spreadsheet. Before applying data-mining techniques, the data were first checked out for questionable data, and those that were found to be unrealistic were cleaned up. Changes made to the data fall under the following categories: deletion of invariant columns, deletion of descriptive columns or columns with so much variety, deletion of redundant columns, deletion of unimportant data and categorization of some columns. After preprocessing the traffic accidents data, the 48 different attributes were reduced to 16 attributes that cover accident, driver, and road/vehicle conditions. Table 2 displays the relevant attributes and their description. Sixteen variables were used with the class variable of degree of injury in an attempt to identify the important variables affecting injury severity of traffic accidents. The data contained information related to accidents, drivers and road conditions. Individual accident records include information about the accident (e.g. year, day, time, accident reason and accident type) and the driver involved (e.g. age, gender, nationality, injury severity level, seat belt). Roadway data

include information on road lighting, road surface condition (e.g. dry, wet, sandy, oily), road speed limit and number of lanes. Weather data include the weather conditions prevailing at the time of the accident such as clear, rain or fog.

Table 2: The 16 selected Attributes

	Attributes Name	Description
Accident Attributes	Year	The year of accident
	Day	The day of accident
	Time	The accident occurred on what time of the day.
	Reason	Reason of the accident
	Accident Type	Type of the accident
Driver Attributes	Gender	Gender of the driver
	Nationality	Nationality of the driver
	Age Rank	Age of the driver
	Seat Belt	The usage of the belt during driving
	Causality Status	Whether the causality is driver, passenger, or pedestrian
	Degree of Injury	Death, sever, moderate, minor.
Road Condition	Lighting	Lighting condition of the road at the time of accident
	Road Surface	Whether the surface of the road was dry, wet, sandy, or oily.
	Speed Limit	The road speed limit
	Lane Numbers	The number of road lanes
	Weather	The condition of the weather

2.2 Random Forests

Standard decision trees are tree-like models in which each internal node represents one of the input variables, and it has a number of branches equal to the number of possible values of that input variable. Each leaf node holds a value of the target attribute. The path from root to leaf represents classification rules. Decision trees begins with the original set *S* as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set *S* and calculates the entropy or information gain of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set *S* is then split by the selected attribute to produce subsets of the data. The algorithm continues to recur on each subset, considering only attributes never selected before. The standard decision trees suffer from the problem of overfitting to the training data. In order to overcome this problem, the splitting at each node is done in RF using the best among a subset of predictors randomly chosen at that node. The RF algorithm works as follows:

1. Draw ***n*** samples from the original data set
2. For each sample, generate a decision tree, with the following modification: at each node, randomly sample ***y*** of the trees and choose the best split from among the variables, where ***y*** is less than ***n***.
3. Classify new data by aggregating the prediction of the ***y*** trees.

The performance of a classifier model is defined from a matrix, known as confusion matrix, which shows the correctly and incorrectly classified instances for each class. Table 3 shows the 2 × 2 confusion matrix for a binary classifier that has only two classes-positive and negative (in our case it becomes 4×4 as we have 4 classes).

Table 3: Confusion Matrix

	Predicated Class	
True Class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The TP, TN, FP, FN can be described as follow:

- True Positive (TP): instances that are positive and classified as positive
- True Negative (TN): instances that are negative but classified as negative
- False Positive (FP): instances that are negative but classified as positive
- False Negative (FN): instances that are positive and classified as negative.

The measures that are used to evaluate the performance of a classifier are computed from the generated confusion matrix. The most widely used evaluation measure is the accuracy rate, which shows the percentage of correctly classified instances and calculated as follow:

$$Accuracy = \frac{TP+Tn}{TP+FP+FN+TN} \quad (1)$$

The predication models built in this paper were built using 10-fold cross validation method. In this method, the sample is randomly split into 10 equal sized subsamples. Of the 10 subsamples, a single subsample is used for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds can then be averaged to produce a single estimation. The number of trees used to grow the forest was 100. The confusion matrix generated by the RF technique is presented in Table 3.

Table 3: The confusion matric generated by RF using the original data set

	Death	**Minor**	**Moderate**	**Severe**	**Accuracy**
Death	13	100	47	6	7.8%
Minor	10	2930	318	31	89%
Moderate	14	820	816	31	48.5%
Severe	2	231	124	27	7%

The accuracy of RF in predicting the severity of road accidents was 68.5%. As presented in Table 3, the model achieves the highest accuracy for the minor class (i.e., 89%). For the severe and death classes, the performance is very poor. This is actually due to the fact these classes are underrepresented in the data set. In order to overcome this problem we created a balanced dataset by applying the Synthetic Minority Oversampling Technique (SMOTE). This technique was applied to make the number of accidents with death and severe classes equal to the number of accidents with moderate class. In order to create a uniform distribution, the number of accidents with minor injuries was reduced using spread subsample technique.

This created a balanced data set with each class represented by 1660 records. The RF technique was then applied on the new data set. The new confusion matrix is presented in Table 4. The predication accuracy of the new model was 78.19%.

Table 3: The confusion matric generated by RF using the balanced data set

	Death	**Minor**	**Moderate**	**Severe**	**Accuracy**
Death	1558	34	52	16	93.8%
Minor	50	1092	422	96	65.7%
Moderate	64	414	1053	129	63.4%
Severe	6	60	105	1489	89.6%

2.2 Ordered Probit Model

Ordered probit is one of the most widely used methods in predicting the outcome of an ordinal dependent variable. It is used when the dependent variable has more than two outcomes, where these outcomes can be ordered. This method is considered state of the art in predicting road accident severity. In this work, the dependent variable (i.e. degree of injury) was transformed from ordinal to numerical (1, 2, 3, 4; for minor, moderate, sever, death, respectively). The R tool was used to perform the ordered probit. The Polr function in the MASS library was used as follows:

PolrFit <- polr(Degrre_Of_Injury ~year + reason + weather + road_surface + accident_type + seat_belt + status + gender + nationality + day + time + age +speed_limit, method="probit", data = data). (2)

The accuracy of ordered probit using the original data set was 59%. The ordered probit was used again to build a prediction model using the balanced data set. The accuracy was found to be 38%

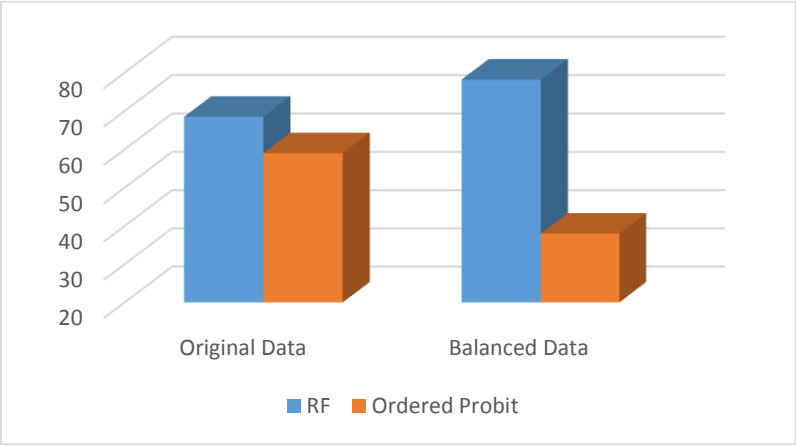

Fig. 2. The performance of RF and ordered probit methods using both original and balanced data set.

3 Discussion and conclusion

In this paper an attempt was made to evaluate the performance of RF technique in predicting the severity or road traffic accidents. Data used to build the prediction models was obtained from Abu Dhabi Emirate for a six-year period (from 2008 to 2013). Using the original data set, the accuracy of the prediction model built using RF was 68.5%, while the accuracy of model built using the ordered probit method was 59%. As shown in Fig.2, It is obvious that RF outperforms ordered probit in predicting the severity of road traffic accidents. Using the balanced data set, RF achieved 14% improvement with prediction accuracy equals to 78.19, while the ordered probit method gave worse prediction accuracy (i.e., 38%).

In summary, it can be concluded that RF technique can be used to effectively predict the severity of road traffic accident. Additionally, it was found that RF gives batter performance than the ordered probit method. Finally, the proportional distribution of the severity classes was found to play a significant role in the achieved prediction accuracy.

References

1. Baluni, Pragya, and Y. P. Raiwani. "Vehicular accident analysis using neural network." International Journal of Emerging Technology and Advanced Engineering 4.9 (2014): 161-164.
2. Shankar, Venkataraman, Fred Mannering, and Woodrow Barfield. "Statistical analysis of accident severity on rural freeways." Accident Analysis & Prevention 28.3 (1996): 391-401.
3. Alkheder, Sharaf, Madhar Taamneh, and Salah Taamneh. "Severity prediction of traffic accident using an artificial neural network." Journal of Forecasting 36.1 (2017): 100-108.
4. Taamneh, Madhar, Salah Taamneh, and Sharaf Alkheder. "Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks." International journal of injury control and safety promotion 24.3 (2017): 388-395.
5. Chang, Li-Yen, and Hsiu-Wen Wang. "Analysis of traffic injury severity: An application of non-parametric classification tree techniques." Accident Analysis & Prevention 38.5 (2006): 1019-1027.
6. Taamneh, Madhar, Sharaf Alkheder, and Salah Taamneh. "Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates." Journal of Transportation Safety & Security 9.2 (2017): 146-166.
7. Kashani, Ali Tavakoli, and Afshin Shariat Mohaymany. "Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models." Safety Science 49.10 (2011): 1314-1320.
8. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
9. Bramer, Max. "Avoiding overfitting of decision trees." Principles of Data Mining. Springer, London, 2013. 121-136.
10. Breiman, Leo. "Manual on setting up, using, and understanding random forests v3. 1." Statistics Department University of California Berkeley, CA, USA 1 (2002).
11. Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.
12. Krishnaveni, S., and M. Hemalatha. "A perspective analysis of traffic accident using data mining techniques." International Journal of Computer Applications 23.7 (2011): 40-48.
13. Das, Abhishek, Mohamed Abdel-Aty, and Anurag Pande. "Using conditional inference forests to identify the factors affecting crash severity on arterial corridors." Journal of safety research 40.4 (2009): 317-327.
14. Haleem, Kirolos, and Albert Gan. "Effect of driver's age and side of impact on crash severity along urban freeways: A mixed logit approach." Journal of safety research 46 (2013): 67-76.
15. Harb, Rami, et al. "Exploring precrash maneuvers using classification trees and random forests." Accident Analysis & Prevention 41.1 (2009): 98-107.
16. Lee, Jaeyoung, BooHyun Nam, and Mohamed Abdel-Aty. "Effects of pavement surface conditions on traffic crash severity." Journal of Transportation Engineering 141.10 (2015): 04015020.