Accident Analysis and Prevention 88 (2016) 37–51

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Bayes classifiers for imbalanced traffic accidents datasets

Randa Oqab Mujalli[a,*], Griselda López[b], Laura Garach[b]

[a] Department of Civil Engineering, The Hashemite University, 13115 Zarqa, Jordan
[b] Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada, Spain

ARTICLE INFO

Article history:
Received 25 April 2015
Received in revised form 24 October 2015
Accepted 2 December 2015

Keywords:
Bayesian networks
Traffic accidents
Urban area
Imbalanced data set
SMOTE

ABSTRACT

Traffic accidents data sets are usually imbalanced, where the number of instances classified under the killed or severe injuries class (minority) is much lower than those classified under the slight injuries class (majority). This, however, supposes a challenging problem for classification algorithms and may cause obtaining a model that well cover the slight injuries instances whereas the killed or severe injuries instances are misclassified frequently. Based on traffic accidents data collected on urban and suburban roads in Jordan for three years (2009–2011); three different data balancing techniques were used: undersampling which removes some instances of the majority class, oversampling which creates new instances of the minority class and a mix technique that combines both. In addition, different Bayes classifiers were compared for the different imbalanced and balanced data sets: Averaged One-Dependence Estimators, Weightily Average One-Dependence Estimators, and Bayesian networks in order to identify factors that affect the severity of an accident. The results indicated that using the balanced data sets, especially those created using oversampling techniques, with Bayesian networks improved classifying a traffic accident according to its severity and reduced the misclassification of killed and severe injuries instances. On the other hand, the following variables were found to contribute to the occurrence of a killed causality or a severe injury in a traffic accident: number of vehicles involved, accident pattern, number of directions, accident type, lighting, surface condition, and speed limit. This work, to the knowledge of the authors, is the first that aims at analyzing historical data records for traffic accidents occurring in Jordan and the first to apply balancing techniques to analyze injury severity of traffic accidents.

1. Introduction

Reducing the severity of accidents is an effective way to improve road safety (Qiu et al., 2014). Recent road traffic safety studies have focused on analysis of risk factors that affect fatality and injury level (severity) of traffic accidents. However, many risk factors are waiting to be discovered or analyzed (Kwon et al., 2015).

Traffic accidents are considered one of the most important and dangerous problems that encounters societies all around the world where it consumes many human and monetary resources. World Health Organization (WHO) statistics indicated that traffic accidents fatalities are estimated to be 1.2 million persons annually worldwide, as well as resulting in 20–50 million injuries. Cost of traffic accidents is estimated to be 518 billion US dollars representing (1–3%) of Gross Domestic Product (GDP) worldwide (WHO, 2013).

Jordan is considered a developing country which has both rapid population and vehicles growth; population statistics of 2013 issued by Department of Statistics (DOS) indicated that Jordan has 6.53 million inhabitants with 1,263,754 registered vehicles (1 vehicle/5 persons) (DOS, 2013). According to Police Traffic Department (PTD) reports for 2013; 107,864 traffic accidents occurred in Jordan with 768 fatalities, 2258 severe injuries and 13,696 slight injuries. A percentage of 94.74% of these accidents were collisions,[1] resulting in 43% of fatalities and 50% of severe injuries (PTD, 2013). Also, 69% of traffic accidents and 71% of collisions occurred in the capital city of Amman, which is considered an urban area having nearly 39% of Jordan's population (2,528,500 inhabitants). In addition, the cost of traffic accidents in Jordan, using unit cost approach to estimate traffic accidents cost in a socioeconomic perspective, is estimated to be 365 million US dollars (PTD, 2013). It is worth noting that Jordan's GDP for 2013 is estimated to be 33.641 billion US dollars, of which the cost of traffic accidents represents 1.2% (DOS, 2013).

* Corresponding author.
E-mail addresses: randao@hu.edu.jo, rmujalli@hotmail.com (R.O. Mujalli).

[1] Collisions exclude all of: run-off-road accidents, pedestrian related accidents and property damage only accidents.

http://dx.doi.org/10.1016/j.aap.2015.12.003
0001-4575/© 2015 Elsevier Ltd. All rights reserved.

Urban and rural accidents characteristics are different (Khorashadi et al., 2005; Theofilatos et al., 2012). Khorashadi et al. (2005) identified significant differences between urban and rural accidents due to differing driver, vehicle, environmental, road geometry and traffic characteristics. Moreover, they estimated that the severe/fatal injury is nearly eight times more likely to occur in an urban area and about 2.5 times more likely in a rural area than other types of injures (i.e. no injury, complaint of pain, or visible injury). Theofilatos et al. (2012) investigated road accident severity with particular focus on the comparison between inside and outside urban areas. They found that factors affecting road accident severity inside urban areas included young driver age, bicyclists, intersections, and collision with fixed objects, whereas factors affecting severity outside urban areas were weather conditions, head-on and side collisions. This demonstrated the particular road users and traffic situations that should be focused on for road safety interventions for the two different types of networks (inside and outside urban areas).

Many modelling techniques have been in use to analyze the injury severity of traffic accidents. The most used models were the logit and probit (Al-Ghamdi, 2002; Milton et al., 2008; Savolainen et al., 2011; Mujalli and De Oña, 2012). However, most of them have their own model assumptions and pre-defined underlying relationships between dependent and independent variables (Chang and Wang, 2006). Recently, many researchers have used methods based on data mining techniques. For example, association rules (Pande and Abdel-Aty, 2009; Montella et al., 2012) or Decision Trees (López et al., 2012a; Abellán el al., 2013; De Oña et al., 2013) have been used for identifying accident patterns. Bayesian networks (BNs) have also been used to study traffic accidents' severity. De Oña et al. (2011) employed BNs to model the relationship between injury severity and variables related to driver, vehicle, roadway, and environment characteristics. They concluded that BNs could be used for classifying traffic accidents according to their injury severity. In addition, Mujalli and De Oña (2011) presented a simplified method based on BNs and variable selection algorithms to predict the injury severity in a traffic accident. Recently, Kwon et al. (2015) used two classification methods, the Naive Bayes and the Decision Tree classifier, for the ranking of risk factors.

Traffic accidents datasets usually have fewer records for fatal and severe injury accidents than for slight injury accidents (Montella et al., 2012). A dataset is considered to be imbalanced if one of the classes (called a minority class) contains a much smaller number of examples than the remaining class (majority class) (Stefanowski and Wilk, 2008). According to Li and Sun (2012) if the proportion of minority class samples constitutes less than 35% of the dataset, the dataset is considered to be imbalanced. Data mining algorithms when learning from imbalanced data tend to produce high predictive accuracy over the majority class, but poor predictive accuracy over the minority class (Thammasiri et al., 2014). Many solutions have been proposed to this problem which can be categorized into two major groups (López et al., 2012b): the internal approaches that create new algorithms or modify existing ones, and the external approaches that preprocess the data in order to diminish the effect of the class imbalance. The pre-processing approach (or resampling techniques) seems to be the more straight forward approach that has greater promise to overcome the class imbalance problem (Thammasiri et al., 2014).

Resampling techniques can be categorized into three groups: the first group consists of the under-sampling methods, which aim to balance the class populations through removing data samples from the majority class until the classes are approximately equally represented. Under-sampling methods randomly eliminate instances from the majority class until a required degree of balance between classes is reached. The second group includes the oversampling methods, which aim to balance class populations through creating new samples from the minority class and adding them to the training set. Finally, the third group comprises the mix methods, which combine both sampling approaches, integrating oversampling of selected minority class instances with removing the most harmful (i.e. noise, and borderline instances that are close to the boundary between the positive and negative classes regions) (Stefanowski and Wilk, 2008; Błaszczyński and Stefanowski, 2015).

In this work, factors affecting injury severity of urban and suburban traffic accidents in Jordan are analyzed. For this purpose, Bayes classifiers are used in the original dataset and in the three balanced datasets (balanced with random under-sampling, with oversampling and with mix methods). Finally, the models developed are compared, and the results for the best model are described.

The paper is organized as follows: Section 2 presents the methodology, the data used, a brief description of Bayes classifiers used, and a description of the performance measures used to evaluate the models. In Section 3, the results and their discussion are presented. Finally, conclusions are given in Section 4.

2. Methodology

In this paper, an imbalanced data set was first obtained and used to develop models applying different popular Bayes classifiers: Efficient Lazy Elimination for Averaged One-Dependence Estimators (AODEsr) (Zheng and Webb, 2006), Weightily Averaged One-Dependence Estimators (WAODE) (Jiang and Zhang, 2006) and Bayesian networks (BNs), where different scores and search algorithms were employed for BNs. Moreover, three balanced datasets were created from the imbalanced data set using three balancing techniques: random under-sampling, oversampling and mix sampling. The same Bayes classifiers used to develop models form the imbalanced data set were also used to develop models from the three balanced data sets. Furthermore, Bayes classifiers were used to analyze injury severity of collisions on urban and suburban roads. The developed models were compared to each other using 10-folds cross validation method, where each data set was first divided into 10 subsets, nine were used to train the model and the remaining one subset was used to test the model. The process was repeated ten times and the average was obtained. As a result, 11 models were developed and compared. Fig. 1 shows the procedure employed.

2.1. Data

Records for traffic accidents which occurred on urban and suburban roads in Jordan were obtained from the Jordanian Police Traffic Department (PTD) for a period of 3 years (2009–2011). The total number of accidents obtained for this period was 49,693. Considering that the main objective of this study was to identify the key factors that contribute to the occurrence of a specific severity in collisions; accidents with property damage only (PDO), Pedestrian and Run-Off-Road were excluded. In this study, only accidents with collisions were analyzed, and as a result, the total number of records used was 16,815.

To identify the main factors that affect urban and suburban collisions severity, fourteen independent variables were analyzed (see Table 1). The variables chosen were based on variables available in the original dataset and the variables used in literature (Theofilatos et al., 2012; Pahukula et al., 2015). The data included variables describing the prevailing conditions at the time of the occurrence of the accident:

- Roadway information: characteristics of the roadway on which the accident occurred such as number of directions, number of lanes, horizontal alignment, grade, pavement type, and pavement surface condition.