Accident Analysis and Prevention 120 (2018) 152–164

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Design and experiment verification of a novel analysis framework for recognition of driver injury patterns: From a multi-class classification perspective

Mengtao Zhu[a], Yunjie Li[a,b,*], Yinhai Wang[b]

[a] School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081, PR China
[b] Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, 98195, USA

ARTICLE INFO

Keywords:
Multi-class imbalanced learning framework
Run-off-road crash
Injury severity pattern
Machine learning
Sensitivity analysis
Traffic safety

ABSTRACT

Detecting driver injury patterns is a typical classification problem. Crash data sets are highly skewed where fatalities and severe injuries are often less represented compared to other events. The severity prediction performance of the existing models is poor due to the highly imbalanced samples of different severity levels within a given dataset. This paper proposes a machine learning based analysis framework from a multi-class classification perspective for accurate recognition of the driver injury patterns. The proposed framework includes preprocessing, classification, evaluation and application of a given dataset. This framework is verified based on the three years single-vehicle ROR (run-off-road) crash records collected in Washington State from 2011 to 2013. At first, thirteen most important safety-related variables are recognized through random forests. Then, the four driver's injury severity levels viz., fatal/serious injury, evident injury, possible injury, and no injury are predicted by integrating the decomposed binary neural network models to achieve better performance. Finally, a sensitivity analysis is carried out to interpret variables' impacts on the decomposed injury severity levels. The study shows that lack of restraint, female drivers, truck usage, driver impairment, driver distraction, vehicle overturn (rollover), dawn/dusk, and overtaking are the leading factors contributing to the driver fatalities or severe injuries in a single-vehicle ROR crash. Most of the findings are consistent with the previous studies. The experimental results validate the effectiveness of the proposed framework which can be further applied for pattern recognition in traffic safety research.

1. Introduction

Investigation of the injury severity patterns in traffic crashes is a fundamental step to develop effective countermeasures for traffic safety improvement, and the prediction of such injury patterns can be viewed as a typical classification problem. Taking the granularity of the crash output variables into consideration, different injury severity levels have been employed in the previous studies. In (Abellán et al., 2013; Zhang et al., 2013 Abu-Zidan and Eid, 2014; Mujalli et al., 2016 Ma et al., 2017 Theofilatos, 2017), the binary injury output levels are considered. On the other hand, several recent studies considered multiple injury severity levels (Kim et al., 2012; Almutairi, 2013; Palamara et al., 2013 Dissanayake and Roy, 2014; Roque et al., 2015 Das and Sun, 2016; Gong et al., 2016 Albdairi and Hernandez, 2017; Gong and Fan, 2017).

The crash datasets usually have much fewer records of fatal and severe injuries compared to other types of accidents as they occur less frequently (Mujalli et al., 2016). When learning from the imbalanced dataset, the traditional classifiers tend to have several problems: (i) These classifiers tend to produce a high accuracy over the majority class but behave badly in minority class (Delen et al., 2006; López et al., 2013; Mujalli et al., 2016). (ii) The training procedure is affected by the considered performance evaluators such as accuracy, which may cause biased learning (Loyola-González et al., 2016). (iii) If the dataset is highly skewed, then the rare samples may be simply treated as noise and vice-versa as discussed in (Beyan and Fisher, 2015). Therefore, the main challenge in the analysis of several injury severity levels is to avoid the imbalanced learning in highly skewed accident observations.

Some efforts have been made to deal with the challenges mentioned above in the crash safety analysis including resampling techniques (Mujalli et al., 2016; Mussone et al., 2017) and feature selection procedure (Mujalli and De, 2011; Yu and Abdelaty, 2013; Chen et al., 2015b Chen et al., 2016a,b; Prati et al., 2017). In (Mujalli et al., 2016),

* Corresponding author at: School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081, PR China.
E-mail address: liyunjie@bit.edu.cn (Y. Li).

https://doi.org/10.1016/j.aap.2018.08.011
Received 5 January 2018; Received in revised form 22 June 2018; Accepted 11 August 2018

the performances of different resampling methods are studied and pointed out that with an imbalanced dataset, the Bayes models are incapable of determining the real influence of the sub-categories "killed" or "severe injury". In (Mujalli and De, 2011), different feature selection methods are analyzed using Bayesian networks and found that a model built with few important variables can produce more useful results. Most of these studies have focused on a selected step in crash analysis without discussing to build a unified analysis framework for imbalanced learning. On the other hand, novel methods such as data mining and machine learning algorithms, which are the newly developed and promising, can be employed for the implementation of the framework for driver injury severity pattern recognition.

In fact, various approaches have been developed in different stages to deal with the imbalanced learning problems in other fields. Several surveys regarding the imbalanced learning have been published (He and Garcia, 2009; Sun et al., 2009 Galar et al., 2012; Fernández et al., 2013; López et al., 2013 Branco and Ribeiro, 2016; Guo et al., 2016). In (Guo et al., 2016), a thorough overview of learning from imbalanced data tasks is provided that include both techniques and applications. It showed the available promising algorithms, plenty of existing applications and future directions. However, the traffic safety-related research is not mentioned. In fact, (Delen et al., 2006) gave a good attempt about imbalanced learning techniques applied to traffic safety research. Their research stressed the imbalanced classification problem and investigated the partitioning methods to measure the variable importance in different decomposed models. Furthermore, this paper proposed a machine learning based analysis framework to deal with the imbalanced learning problem in transportation safety research. This unified framework includes four steps as preprocessing, classification, evaluation and application. Each step can try to apply different analysis methods in different applications. Resampling techniques, feature engineering and cost-sensitive learning (Fernández et al., 2013; Li et al., 2016) can be used in the preprocessing step. Classification step contains binary decomposition, ensemble methods and classifier modifications (Sun et al., 2009; Galar et al., 2012), while the performance metrics like accuracy, Receiver Operating Characteristics (ROC), Area Under Curve (AUC) and F-measure (De et al., 2011; Chen et al., 2015a,b) can be applied in the evaluation step. Compared with (Delen et al., 2006)'s previous research, this paper put more focus on the design and experiment verification of the proposed analysis framework while partitioning methods only played one important processing step in it. At the same time, adoptions of robust variables importance ranking method, ensemble method for integration of the decomposed models results and discussions of variable impacts towards driver injury severity levels contributed more to the differences between this study and Delen's research.

Run-off-road (ROR) crashes have been a major cause of fatalities and serious injuries in the United States. Statistics from the Fatality Analysis Reporting System (FARS) illustrate that the traffic fatalities due to ROR crashes in the United States count up to about 38% of the total traffic fatalities in 2007–2016 (NHTSA, 2017). Since the ROR crashes are more likely to cause either fatalities or severe injuries than the other type of vehicle crashes, it has drawn a lot of attention not only in the United States (Lee and Mannering, 2002; Dissanayake, 2003; Peng and Boyle, 2012; Roy and Dissanayake, 2013; Dissanayake and Roy, 2014; Albdairi and Hernandez, 2017) but also across the world (Palamara et al., 2013; Petegem and Wegman, 2014; Shawky et al., 2014 Roque et al., 2015). The proposed framework is implemented and verified on three years of single vehicle ROR crash records in this study.

The rest of the paper is organized as follows: Section 2 describes the data and preprocessing operations. Model description and implementation specifications are introduced in Section 3. Section 4 discusses the injury pattern from the aspects of variable impact and elasticity. Conclusions and summary of the future direction are presented in Section 5.

2. Data description

This study focuses on three-years single-vehicle ROR crash data collected from 2011 to 2013 in Washington State. The raw data were provided by the Washington State Department of Transportation (WSDOT) which is recording the traffic accident information among all the state routes statewide. The entire dataset consists of five parts regarding the traffic accident: four crash related variables including (i) spatial-temporal information, (ii) environmental information, (iii) driver demographic characteristics, (iv) vehicles related data and one output variable of driver injury severity levels when accidents occurred.

1) The spatial-temporal information describes the crash time, crash location and roadway type.
2) Environmental information consists of weather conditions, road surface conditions, lighting condition and road's geographical features.
3) Driver demographic characteristics refer to driver's age, gender, sobriety degree, behavior, ejection, and restraint usage.
4) Vehicles related data composed of vehicle type, manufacture, usage and movement.
5) Driver injury severity levels record the ROR crash outcomes and are the dependent variables in this study.

The severity criterion of driver injury adopted by most Department of Transportation in different states was KABCO: fatality (K), incapacitating injury (A), visible injury (B), complaint of injury (C), and no apparent injury (O). In imbalanced learning problem, if the proportion of minority class samples is less than 35% of the dataset, then the dataset is considered imbalanced (Li and Sun, 2012). The available ROR dataset was scrutinized and presented in Table 1, and it was found that different driver injury severity levels have an imbalance distribution. Therefore, four categories of injury severity levels are employed in this study as follows: (i) fatality or serious injury(F/S), (ii) evident injury (EI), (iii) possible injury (PI), and (iv) property damage only (PDO).

The observations of ROR crashes were obtained through screening the candidate variables "driver sequence 1", and the events with records "Ran off the road" were selected. A ROR crash dataset including 12,788 single-vehicle ROR accident observations was formed from overall 133,579 accidents. Further, eighteen variables were selected from the original reports according to the previous driver injury severity studies (Huang and Abdel-Aty, 2010; Savolainen et al., 2011 Mannering and Bhat, 2014) and engineering experience, all of which were categorical variables and were coded numerically. Variables composed of multiple attributes and highly skewed in frequency were decomposed, for example, road characteristics were divided into road curvature and road grade. Similarly, for multi-categorical variables, those values with similar contributions were merged, for example, "apparently asleep" and "apparently fatigued" under the variable "driver action" were reduced to "fatigue". From these records, the incomplete and "unknown" crash records were screened out. Such processing is consistent with the existing literature in driver injury severity analysis. A detailed variable definitions and data description are shown in Table 1. The interpretation of variables as well as the numerical coding value for each sub-category within every variable are illustrated, accompanied with driver injury severities corresponding to the value.

3. Methodology

3.1. Multi-class imbalanced learning framework

To gain more insight into the mechanism of accident crashes, a higher prediction performance and a better understanding are essential. This article proposed the multi-class imbalanced learning framework into the analysis of accident crashes.