Accident Analysis and Prevention 111 (2018) 354–363

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network

Yunjie Li[a,b], Dongfang Ma[c,b], Mengtao Zhu[a], Ziqiang Zeng[d,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
[c] Institute of Marine information science and technology, Zhejiang University, Zhoushan 316021, PR China
[d] Uncertainty Decision-Making Laboratory, Sichuan University, Chengdu 610064, PR China

ARTICLE INFO

Keywords:
Significant factor
Highway crash
Genetic algorithm
Neural network
Traffic safety

ABSTRACT

Identification of the significant factors of traffic crashes has been a primary concern of the transportation safety research community for many years. A fatal-injury crash is a comprehensive result influenced by multiple variables involved at the moment of the crash scenario, the main idea of this paper is to explore the process of significant factors identification from a multi-objective optimization (MOP) standpoint. It proposes a data-driven model which combines the Non-dominated Sorting Genetic Algorithm (NSGA-II) with the Neural Network (NN) architecture to efficiently search for optimal solutions. This paper also defines the index of Factor Significance (F_s) for quantitative evaluation of the significance of each factor. Based on a set of three year data of crash records collected from three main interstate highways in the Washington State, the proposed method reveals that the top five significant factors for a better Fatal-injury crash identification are 1) Driver Conduct, 2) Vehicle Action, 3) Roadway Surface Condition, 4) Driver Restraint and 5) Driver Age. The most sensitive factors from a spatiotemporal perspective are the Hour of Day, Most Severe Sobriety, and Roadway Characteristics. The method and results in this paper provide new insights into the injury pattern of highway crashes and may be used to improve the understanding of, prevention of, and other enforcement efforts related to injury crashes in the future.

1. Introduction

According to statistics from the National Highway Traffic Safety Administration (NHTSA, 2017), which stores road data for the 50 US states, in 2015 alone, nationwide, there were a total of 22,441 deaths and 2.18 million injuries due to automobile accidents. Since fatal highway crashes are a major cause of injury, death, and economic loss, the identification of the significant factors associated with such crashes has become a major interest in transportation safety research. One challenge with such studies is that potential redundant information and correlations among different candidate factors must be addressed properly and effectively. As the most common crash data set, the police crash report contains almost all the related variables describing a crash. For example, there are more than one hundred factors in the crash report recorded by the Washington State Police (WSDOT, 2014). The academic community has made continuous efforts to develop more robust and efficient methods for the exploration and analysis of such data.

In literature, statistical models have been the primary method for the analysis of such data. From an early stage, methods such as Logistic regression (Singleton et al., 2004; Dissanayake and Lu, 2002; Hanrahan et al., 2009) have been commonly used for such analysis. Subsequently many researchers have applied novel methods to broaden the scope of applicability of the statistical models. Due to the ordinal nature of the injury outcomes (for example, ranging from no injury, to injury to fatal), ordered choice models have also been popularly applied to the analysis and severity modeling of the crash injury data (Kockelman and Kweon, 2002; Kaplan and Prato, 2012; Mohamed et al., 2013). More recently, to take into account the limitation of the assumption that all parameters estimated in the models were constant across observations and to address the heterogeneity of the crash outcomes, some multinomial logit models (Hu et al., 2010; Hu and Donnell, 2011; Shankar and Mannering, 1996) and mixed logit models (Milton et al., 2008; Malyshkina and Mannering, 2010; Zeng et al., 2017) have been developed to analyze the crash injury severities. However, the mass of complicated data on crashes nowadays will still make it difficult to use statistical models to investigate the factors related to injury severity efficiently. One restriction of such analysis is the requirement that the

* Corresponding author.
E-mail address: yinhai@edu.edu (Y. Wang).

https://doi.org/10.1016/j.aap.2017.11.028
Received 19 September 2017; Received in revised form 20 October 2017; Accepted 21 November 2017

data must meet some statistical assumptions (Harrell, 2001; Cohen et al., 2003; Tabachnick and Fidell, 2012) while such assumptions are hard to be valid in most crash circumstances. Another drawback of such an approach is due to their poor performance in handling several discrete variables or variables with a high number of categories (Cohen et al., 2003; Tabachnick and Fidell, 2012).

To overcome the shortcomings of statistical models, researchers have proposed many non-parametric models and artificial intelligence models for the study of crash injury patterns. The Classification and Regression Tree (CART), is a non-parametric model without any predefined underlying relationship between the dependent and independent variables. It has been widely employed for the study of crash outcomes (Chang and Wang, 2006; Yan and Radwan, 2006; Pande and Abdel-Aty, 2006; Chen et al., 2016a). The Support Vector Machine (SVM) model is also a relatively new method to solve classification problems, and has been utilized for the classification of crash injury severity (Li et al., 2012; Yu and Abdel-Aty, 2014; Chen et al., 2016b). Li et al. applied the SVM model for crash injury severity analyses, concluding that SVM models outperform the popular ordered probit model for the prediction of injury severity and factor impact assessment (Li et al., 2012). Yu and Abdel Aty compared the performance of the SVM model, random parameter models, and fixed parameter models to predict the severity of crash injuries. They concluded that SVM and random parameter models outperform fixed parameter models (Yu and Abdel-Aty, 2014). Artificial neural network (ANN), have also been applied for a long time for the classification of crash severity analysis; and recently the applications of this method have grown immensely (Abdelwahab and Abdel-Aty, 2003; Lu et al., 2012; Ali and Tayfour, 2012). All these methods have demonstrated powerful and adaptive analysis capabilities, and researchers have drawn several useful conclusions using these methods. With the coming of a big-data era, optimization efficiency has become a primary concern during the analysis procedure. For the identification of significant factors related to crash severity, performance of CART models is highly dependent on the values of the parameters and the generated model ends up with a weak generalization capability (Harrell, 2001; Chang and Chen, 2005). Also, the use of a greedy searching strategy in the CART method requires an exhaustive search procedure which can be very time consuming (Kashani et al., 2011; Prati et al., 2017). For SVMs and ANNs, the models themselves lack the capability of automatically selecting significant factors contributing to the target variable (Chen et al., 2016b).

To summarize, analysis methods used in previous studies still face a few challenges or have a limited efficiency. There is a need to develop novel algorithms to efficiently handle the traffic crash records. Considering the process of identification of significant factors as a multi-objective optimization (MOP) problem, genetic algorithm techniques may be applied as new optimal factors searching algorithms to improve the performance of the analysis (Li et al., 2012). The Non-dominated Sorting Genetic Algorithm (NSGA-II) (Deb et al., 2002) is a fast elitist multi-objective genetic algorithm. It has already found many applications in different fields, such as spectrum assignment in a spectrum sharing networks (Martínez-Vargas et al., 2016), modeling and control of output fiber length distribution in paper-making (Zhou et al., 2017), improvement of dynamic cellular manufacturing systems (Azadeh et al., 2017), and traffic signal optimization (Branke et al., 2007). However, to our knowledge, no research has used the NSGA-II algorithm for the study of injury patterns in highway crashes. The main aim of this paper is to propose a new hybrid model combining the NSGA-II algorithm and the Neural Network (NN) model for significant factors identification in highway fatal-injury crashes. Results of the study have demonstrated its capability for the searching and evaluation of optimal solutions in the global or sub-global space. The rest of this paper is organized as follows: the data description and preprocessing procedure are provided in Section 2. Sections 3 and 4 present the NSGA-NN hybrid model and discuss the analysis results. Finally, in Section 5 we present a summary about the findings.

Table 1
Summary statistics of crash data for three highways in Washington between 2011 and 2013.

		Total Crashes	Fatality-Injury Crashes		PDO Crashes	
Data in Different Year	D2011	11621	3669	31.6%	7952	68.4%
	D2012	11878	3654	30.8%	8224	69.2%
	D2013	11670	3302	28.3%	8368	71.7%
Data in Different Route	D-I5	23264	6869	29.5%	16395	70.5%
	D-I90	5988	1897	31.7%	4091	68.3%
	D-I405	5917	1859	31.4%	4058	68.6%

2. Data description

This study was performed with data from police crash reports provided by the Washington State Department of Transportation (WSDOT). The data consists of records of three years from January 2011 to December 2013 for three main Interstate highways (including I-5, I-90, and I-405), collected in the Washington State (Table 1). The total number of crashes during this period was 35,169. All the data are divided into three sets, by year, denoted as D2011, D2012 and D2013 from a temporal perspective. Additionally, they are divided from a spatial perspective (i.e., by interstate highway), into three other sets called D-I5, D-I90 and D-I405.

The severity of a crash severities is usually the target variable for the analysis of most injury patterns. According to the WSDOT, crash severity levels are often referred to using the KABCO scale, which uses the parameters: fatal (K), incapacitating-injury (A), non-incapacitating injury (B), minor injury (C), and property damage only (PDO or O)). Most of previous literatures aggregated five KABCO levels into 3 levels or 2 levels. Different divisions of two level include PDO and Injury/Fatal (Ma et al., 2017), Minor Injury and Severe/Fatal (Theofilatos, 2017; Abellán et al., 2013; Zhang et al., 2013; Mujalli et al., 2016), Death and not Death (Abu-Zidan and Eid, 2014). This study regards people are the most important part to be protected in an accident even they suffer minor injuries only, so it have broken down all the collected crash data into two categories labeled as Fatal-injury crashes (including KABC) and PDO crashes.

Over one hundred items describing the characteristics of a crash are included in the police crash report in Washington. Weather/Time, Road, Vehicle, and Driver are four categories of such items that may affect traffic safety. Based on the strong correlation among the factors, this study selects 14 factors which fall into one of these four categories. Before discussing the methodology, we describe the data cleaning and preprocessing steps. First, all incomplete records containing an “unknown” or “none” field were eliminated from the original data set. Next, for those factors that include many detailed values, a combination operation was applied to combine these small values into a single new larger value. The principle behind the combination of these different values mainly focuses on the similarity of their effects on the traffic safety. For example, “Driver Conduct” has 35 different values in the original report. “Apparently Asleep”, “Apparently Fatigued,” and “Apparently Ill” were combined into a new grouped value since they all represent cases of driver ailment. Another operation was converting the numerical factors into enumerated values. For example, all the twenty four hours in a day of the Hour of Day field were replaced with four categorical values as “midnight,” “morning rush hours,” “daytime” and “afternoon rush hours”. Table 2 shows the grouping and summaries of the value definitions for all the selected candidate factors.

3. Methodology

Fig. 1 presents the research outline of this paper. The methodology consists of two parts: a) A factors optimization model constructed on the multi-objective optimization (MOP) idea and b) a quantifying index