Accident Analysis and Prevention 75 (2015) 1–15

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Application of classification algorithms for analysis of road safety risk factor dependencies

Oh Hoon Kwon [a], Wonjong Rhee [b], Yoonjin Yoon [a,*]

[a] Department of Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, South Korea
[b] Graduate School of Convergence Science and Technology, Seoul National University, 145 Gwanggyo-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 443-270, South Korea

ARTICLE INFO

Article history:
Received 18 July 2014
Received in revised form 13 October 2014
Accepted 1 November 2014
Available online 16 November 2014

Keywords:
Traffic accident analysis
Severity level
Risk factor
Naive Bayes classifier
Decision tree classifier
Dependency

ABSTRACT

Transportation continues to be an integral part of modern life, and the importance of road traffic safety cannot be overstated. Consequently, recent road traffic safety studies have focused on analysis of risk factors that impact fatality and injury level (severity) of traffic accidents. While some of the risk factors, such as drug use and drinking, are widely known to affect severity, an accurate modeling of their influences is still an open research topic. Furthermore, there are innumerable risk factors that are waiting to be discovered or analyzed. A promising approach is to investigate historical traffic accident data that have been collected in the past decades. This study inspects traffic accident reports that have been accumulated by the California Highway Patrol (CHP) since 1973 for which each accident report contains around 100 data fields. Among them, we investigate 25 fields between 2004 and 2010 that are most relevant to car accidents. Using two classification methods, the Naive Bayes classifier and the decision tree classifier, the relative importance of the data fields, i.e., risk factors, is revealed with respect to the resulting severity level. Performances of the classifiers are compared to each other and a binary logistic regression model is used as the basis for the comparisons. Some of the high-ranking risk factors are found to be strongly dependent on each other, and their incremental gains on estimating or modeling severity level are evaluated quantitatively. The analysis shows that only a handful of the risk factors in the data dominate the severity level and that dependency among the top risk factors is an imperative trait to consider for an accurate analysis.

1. Introduction

As human society continues to become more complex, the role of transportation continues to become more essential for human society. Transportation has literally converted impossible ideas into amazing realities. Transportation, however, comes with accidents that create mild to fatal injuries, and reducing the fatalities and injuries of road traffic accidents, has long been a concern for governments and the general public. Countless studies have been performed to understand the risk factors that contribute to traffic accidents. An example is the World Health Organization (2004) that analyzes the risk factors for injury severity and injury frequency. In most of the existing studies of risk factors, the data are categorized according to the accident severity level, for instance, fatal, severe, slight injury, and property damage only. The other data fields can be used to model or estimate the severity level, and the modeling or estimation allows a common and effective way of evaluating risk factors and their impact (Savolainen et al., 2011).

Existing studies have adopted a variety of methodologies for risk factor analysis. Regression models such as logit and probit have been widely employed (Al-Ghamdi, 2002; Kockelman and Kweon, 2002; Sze and Wong, 2007). In regression models, binary or multiple levels of severity are typically set as dependent variables, and the risk factors that can affect severity are set as independent variables. A common assumption for regression models is that there is no dependency among the risk factors. In addition to the 'no dependency' assumption, regression models need to assume a specific functional form to model the relationships between dependent and independent variables. Thus, regression models can be limited if the assumptions do not hold well (Chang and Wang, 2006).

In order to overcome the limitations of regression models, classification models of data-mining approaches have been

[*] Corresponding author. Tel.: +82 42 3503615; fax: +82 42 3503610.
E-mail addresses: ohunkwon@kaist.ac.kr (O.H. Kwon), wrhee@snu.ac.kr (W. Rhee), yoonjin@kaist.ac.kr (Y. Yoon).

http://dx.doi.org/10.1016/j.aap.2014.11.005
0001-4575/© 2014 Elsevier Ltd. All rights reserved.

applied to the risk factor analysis problem. A classifier is a function that classifies the class variable given a set of input variables which are called feature or attribute variables. Typically, severity level is set as a class variable and risk factors are set as feature variables (Sohn and Shin, 2001; Chang and Wang, 2006; Beshah and Hill, 2010; Kashani and Mohaymany, 2011,b; Montella et al., 2011a,b; Shanthi and Ramani, 2012). Among the classification models, the Bayesian classifier and the decision tree classifier are commonly used in data mining. The Bayesian classifier is a statistical model based on Bayes' rule; the model can explain dependencies among the risk factors. The Bayesian classifier, however, can be computationally expensive, especially when the data have a large dimension (Heckerman et al., 2001), because it needs to construct a Bayesian network that is a directed, acyclic graphical model. Instead, the Naive Bayes classifier, also called the simple Bayes classifier, allows a computationally inexpensive learning of the model, a conspicuous interpretation, and an accurate classification under the assumption that feature variables are conditionally independent of each other for a given class variable (Xhemali et al., 2009). In other words, if the risk factors for a given severity level (e.g., fatal) are independent, the Naive Bayes classifier is known to work well. The decision tree does not require any specific functional form to build a model, and the model can be easily interpreted for obtaining insights on risk factors. Furthermore, the decision tree does not require any assumption on dependency among the risk factors, and it is known to work well regardless of the dependency among the data fields (Beshah and Hill, 2010).

In this study, two classification methods, the Naive Bayes and the decision tree, are used to model the severity of traffic accidents using available data fields, i.e., risk factors. A typical binary logistic regression model is additionally used as the basis for comparisons of their performances. While most of the previous studies have focused on the identification itself of important risk factors, this study focuses on the dependency among the risk factors. Using the fact that the Naive Bayes does not consider dependency while the decision tree does, a few novel methodologies are employed to analyze the dependency among the most important risk factors. Furthermore, the impact of dependency is evaluated in a quantitative way by plotting receiver operating characteristics (ROC) and precision–recall (PR) curves. To double confirm the results, an interesting methodology is applied where the top ranking risk factors from the Naive Bayes and the top ranking risk factors from the decision tree are respectively used as the feature variables of the Naive Bayes. To be more specific, performance of the Naive Bayes is compared for the two different sets of feature variables.

The data used for this study are the statewide integrated traffic records system (SWITRS) data, which are a collection of traffic accident reports accumulated by the California Highway Patrol (CHP) since 1973. Among various vehicle/pedestrian groups, this study focuses on accidents involving passenger cars, sport utility vehicles (SUVs), and minivans from 2004 to 2010. Even though the original data provide five different levels of severity, this study uses two severity levels (fatal or injured vs. property damage only) so that the classifiers are not affected by the bias in the sample size per severity group.

In Section 2, we review earlier studies on analysis of traffic accident severity. In the following section, we describe the accident data used in this study and present the data fields that are used as feature and class variables in the classifiers. In Section 4, we explain the methodology of the Naive Bayes and decision tree classifiers and the ranking of the risk factors. In Sections 5 and 6, the results of the analysis are displayed and we discuss them. Finally, we conclude this study in Section 7.

2. Literature review

Existing studies on analyzing the injury severity of accidents used a variety of methodological approaches, logit and probit regression models for categorical data being the most common. In Al-Ghamdi (2002), logistic regression was employed with a binary dependent variable (fatal and non-fatal) and nine independent variables. The model identified the two most significant factors: location and the cause of the accident. Sze and Wong (2007) used a binary logistic regression model to analyze fatalities and severe injuries of pedestrians. They introduced a temporal variable into the model to consider temporal change in pedestrian injuries. One of the major findings of the study was that controlling the significant factors tended to decrease the risk of pedestrian injury. Kockelman and Kweon (2002) applied ordered probit models to three categories of crash type: all crash, two-vehicle crash, and single-vehicle crash. Considering four injury severity levels as the dependent variable, the study concluded that passenger cars were safer than pickup trucks and SUVs in a single-vehicle crash. In a two-vehicle crash, pickup trucks and SUVs tended to protect their drivers better but increased injury severity for occupants of the other vehicles involved. Savolainen et al. (2011) summarized other existing research using various logit and probit regression models. The study discussed methodological challenges, such as addressing the spatial and temporal correlation, under-reporting of accidents, and new types of accident data.

With the increasing availability and accessibility of a large amount of historical accident data, numerous recent studies adopted data mining techniques. The decision tree has been used to identify important factors and patterns of accident severity. Chang and Wang (2006) used the classification and regression tree (CART) model to establish the relationship between injury severity and influential factors. In the study, the authors assumed the injury severity level of an accident is the injury level of the worst-injured occupant and categorized it into three levels: fatality, injury, and no-injury. They concluded that the vehicle type is the most influential factor in injury severity. Kashani and Mohaymany (2011) presented a study on the injury severity of vehicle passengers on two-lane, two-way rural roads in Iran. The CART model was employed, and binary injury severity of the occupants was used to improve classification accuracy from an imbalance class problem. The study found that two most important factors were improper overtaking and not using a seat belt. Montella et al. (2011a,b); Montella et al. (2011a,b) used the decision tree and association rules to analyze accidents involving pedestrians and powered two-wheeler vehicles. For pedestrians, the most influential factors were road type, pedestrian age, lighting conditions, and vehicle type. For powered two-wheeler vehicles, the curve alignment of the road, rural area, run-off-the-road crash, night time, rainy weather, and greater cylinder capacity of the vehicle were significantly associated with fatal severity. In addition, the authors analyzed dependencies among the factors by setting the most significant factors as class variables in each classifier. Based on the results, they concluded that there were dependencies among the factors and that results of the rule-based analysis were consistent with the results of other existing studies using probabilistic models.

Existing studies employing classification models for severity analysis of traffic accidents have generally used single-number measures such as accuracy, precision, and recall to evaluate the models and compare the performances of several classification models. Sohn and Shin (2001) compared neural network, logistic regression, and decision tree using accuracy measure, and found that the three models exhibited a similar level of classification accuracy. Shanthi and Ramani (2012) applied various popular decision tree algorithms, such as ID3, C4.5, and the CART and Naive