Transportation Research Part C xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Transportation Research Part C

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

A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction

Lei Lin, Qian Wang, Adel W. Sadek *

Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, USA

ARTICLE INFO

Article history:
Received 15 October 2014
Received in revised form 9 March 2015
Accepted 10 March 2015
Available online xxxx

Keywords:
Frequent Pattern tree (FP tree)
Fuzzy C-means clustering (FCM)
Bayesian network
k Nearest Neighbor (*k*-NN)
Variable importance
Variable selection
Random forest
Real time
Relative Object Purity Ratio (ROPR)
Traffic accident risk prediction

ABSTRACT

With the availability of large volumes of real-time traffic flow data along with traffic accident information, there is a renewed interest in the development of models for the real-time prediction of traffic accident risk. One challenge, however, is that the available data are usually complex, noisy, and even misleading. This raises the question of how to select the most important explanatory variables to achieve an acceptable level of accuracy for real-time traffic accident risk prediction. To address this, the present paper proposes a novel Frequent Pattern tree (FP tree) based variable selection method. The method works by first identifying all the frequent patterns in the traffic accident dataset. Next, for each frequent pattern, we introduce a new metric, herein referred to as the Relative Object Purity Ratio (ROPR). The ROPR is then used to calculate the importance score of each explanatory variable which in turn can be used for ranking and selecting the variables that contribute most to explaining the accident patterns. To demonstrate the advantages of the proposed variable selection method, the study develops two traffic accident risk prediction models, based on accident data collected on interstate highway I-64 in Virginia, namely a *k*-nearest neighbor model and a Bayesian network. Prior to model development, two variable selection methods are utilized: (1) the FP tree based method proposed in this paper; and (2) the random forest method, a widely used variable selection method, which is used as the base case for comparison. The results show that the FP tree based accident risk prediction models perform better than the random forest based models, regardless of the type of prediction models (i.e. *k*-nearest neighbor or Bayesian network), the settings of their parameters, and the types of datasets used for model training and testing. The best model found is a FP tree based Bayesian network model that can predict 61.11% of accidents while having a false alarm rate of 38.16%. These results compare very favorably with other accident prediction models reported in the literature.

1. Introduction

Traffic accidents cause a great deal of loss of lives and property. According to the accidents report of the United States Census Bureau, there were 10.8 million accidents and 35,900 persons killed in 2009 (US census bureau, 2013). To address this, many studies have been conducted to predict accident frequencies and analyze the characteristics of traffic accidents, including studies on hazardous location/hot spot identification (Lin et al., 2014), accident injury-severities analysis (Milton et al., 2008), and accident duration analysis (Zhan et al., 2011).

* Corresponding author. Tel.: +1 716 645 4367x2421; fax: +1 716 645 4367.
E-mail address: asadek@buffalo.edu (A.W. Sadek).

http://dx.doi.org/10.1016/j.trc.2015.03.015
0968-090X/

With the development of intelligent transportation systems technologies, there currently exists a wealth of real-time traffic data collected from fixed-locations sensors, automatic vehicle identification systems and other sensing technologies. These data sources can be fused and analyzed to develop real-time management strategies and applications for the purpose of improving efficiency, safety, resiliency and reliability of transportation systems. Particularly in the area of transportation safety, researchers have started to develop real-time traffic accident risk prediction models that take advantage of complex and rapidly and continuously flowing data for predicting traffic accidents.

New issues are emerging accompanying the new opportunities offered by real-time traffic data. One issue is that related to explanatory variable selection, a topic that has received increased attention in real-time traffic accident risk prediction. The wealth of real-time traffic data offer more explanatory variables that may contribute to explaining traffic accident risk and patterns. However, as has been widely recognized, "more is not always better", particularly for accident prediction. Inclusion of a large number of explanatory variables may cause model overfitting (Sawalha and Sayed, 2006). In addition, it can cause application related issues such as long prediction running time and unreliable prediction results, particularly when a model is applied to new locations and larger data instances (Fernández et al., 2014).

In terms of usage, as a preprocessing step before building any prediction models, variable selection can help researchers identify and extract meaningful information (patterns, structure, underlying relationships, etc.) from the data. Only a small representative subset of the original feature space of the data may be needed to interpret the results (Fernández et al., 2014).

Real-time traffic accident risk prediction models can be broadly classified into two categories, namely statistical models and data mining/machine learning models. Statistical models, such as matched case-control logistic regression models (Abdel-Aty et al., 2004), binary logit models (Xu et al., 2013) and aggregate log-linear models (Lee et al., 2003), have been tested and used in the previous studies. Typical examples of the data mining/machine learning modeling approach include k nearest neighbor models (Lv et al., 2009), neural networks (Abdel-Aty et al., 2008), Bayesian network models (Hossain and Muromachi, 2012) and support vector machines (Yu and Abdel-Aty, 2013). Those methods have been gaining more and more popularity in recent years.

As previously mentioned, the variable selection problem has attracted attention in previous real-time traffic accident risk prediction research. For statistical models, Sawalha and Sayed (2006) found that using less but statistically significant explanatory variables can avoid over-fitting and improve the reliability of a model. They suggested combining the *t*-statistics test and the likelihood ratio based scaled deviance test, for selecting significant explanatory variables. Different procedures were suggested for Poisson regression and negative binomial regression respectively due to the additional complexity introduced to the scaled deviance test for negative binomial regression models. As for the data mining models, classification and regression tree (CART) has been used to perform variable selection (Yu and Abdel-Aty, 2013; Pande and Abdel-Aty, 2006). Another ensemble learning method for classification and regression, called random forest, has also been widely used to rank explanatory variables (Abdel-Aty et al., 2008; Ahmed and Abdel-Aty, 2012). Recently, a hybrid model random multinomial logit (RMNL), formed by combining the random forest and logit models, was applied to calculate traffic accidents variable importance (Hossain and Muromachi, 2012).

Different from previous studies, this paper proposes a novel frequent pattern tree (FP tree) based variable selection method for real-time traffic accident risk prediction, using the data collected on interstate highway I-64 in Virginia as the case study. A new algorithm was built to rank explanatory variables based on the "calculated variable importance score". To verify the model performance, the study then develops two traffic accident risk prediction models, namely a *k*-nearest neighbor model and a Bayesian network model. Prior to the model development, two variable selection methods are utilized: (1) the FP tree based method proposed by the present research; and (2) the baseline random forest tree based method. The results show that the models trained with the FP tree selected explanatory variables always outperformed the others regardless of the types of the prediction models, their parameter settings, and the types of datasets used for model training and testing. To the best of the authors' knowledge, this paper is the first attempt toward applying the FP tree based models to traffic accident related research. It is also one of the few studies that focus on the real-time prediction of traffic accident risk.

The paper is organized as below. First, in the methodology section, we introduce the FP tree model and its variable importance score calculation algorithm. Second, we describe the traffic accident datasets used for model training and testing. Third, prior to the risk prediction model development, we describe and compare the FP tree and the random forest based variable selection methods in terms of their variable importance ranking results. Fourth, based on the variables selected by the FP tree and the random forest methods respectively, two traffic accident risk predictions models are discussed and compared in terms of their prediction performance, namely the *k*-NN model and the Bayesian network model. The paper ends with a summary of the main conclusions of the work and suggestions for future research.

2. Model methodology

This section discusses the FP-tree algorithm used in this paper for explanatory variable selection. The algorithm consists of two steps: variable discretization and variable importance score calculation. For the former step, the fuzzy c-means clustering method is used to convert a continuous variable to a series of discrete categorical variables; for the latter, we propose the "Relative Object Purity Ratio (ROPR)" as an importance score for each explanatory variable. This section will also introduce the random forest method that is used as the bench-marking variable selection method. Finally, the two methods used for accident risk prediction, namely the *k*-NN model and Bayesian network, are briefly introduced.

Please cite this article in press as: Lin, L., et al. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.03.015