Accident Analysis and Prevention 82 (2015) 192–198

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects

Ni Dong, Helai Huang*, Liang Zheng

Urban Transport Research Center, School of Traffic and Transportation Engineering, Central South University, Changsha, Hunan, 410075 PR China

ARTICLE INFO

Article history:
Received 15 December 2014
Received in revised form 22 May 2015
Accepted 26 May 2015
Available online 16 June 2015

Keywords:
Support vector machine
Spatial weight features
CAR model
Correlation-based feature selector

ABSTRACT

In zone-level crash prediction, accounting for spatial dependence has become an extensively studied topic. This study proposes Support Vector Machine (SVM) model to address complex, large and multi-dimensional spatial data in crash prediction. Correlation-based Feature Selector (CFS) was applied to evaluate candidate factors possibly related to zonal crash frequency in handling high-dimension spatial data. To demonstrate the proposed approaches and to compare them with the Bayesian spatial model with conditional autoregressive prior (i.e., CAR), a dataset in Hillsborough county of Florida was employed. The results showed that SVM models accounting for spatial proximity outperform the non-spatial model in terms of model fitting and predictive performance, which indicates the reasonableness of considering cross-zonal spatial correlations. The best model predictive capability, relatively, is associated with the model considering proximity of the centroid distance by choosing the RBF kernel and setting the 10% of the whole dataset as the testing data, which further exhibits SVM models' capacity for addressing comparatively complex spatial data in regional crash prediction modeling. Moreover, SVM models exhibit the better goodness-of-fit compared with CAR models when utilizing the whole dataset as the samples. A sensitivity analysis of the centroid-distance-based spatial SVM models was conducted to capture the impacts of explanatory variables on the mean predicted probabilities for crash occurrence. While the results conform to the coefficient estimation in the CAR models, which supports the employment of the SVM model as an alternative in regional safety modeling.

1. Introduction

Crash prediction model (CPM) is an essential tool in traffic safety analysis. Numerous applications have been developed to evaluate safety level of various types of road entities and to examine effect of safety countermeasures. Recently, traffic crashes are aggregated by a certain spatial scale and researchers usually seek to relate safety to zone-level factors. One of the main objectives of macro-level crash prediction analysis is to explain observed cross-sectional variations in safety using zone-level covariates at different spatial scales (e.g., states, counties, traffic analysis zones, and census wards) (Washington et al., 2006; Quddus 2008; Huang et al., 2010). These macro-level CPMs may aid transportation agencies in more effectively incorporating safety consideration into transportation planning and management (Abdel-Aty et al., 2011; Huang et al., 2013; Xu and Huang, 2015).

In zonal crash prediction, accounting for spatial dependency has become an extensively studied topic. Previous studies (i.e., Quddus 2008; Huang and Abdel-Aty 2010; Siddiqui et al., 2012; Xu et al., 2014; Zeng and Huang, 2014a) found that traffic crashes exhibit extensive spatial dependency across neighboring zones. Research commonly seeks to address the issue of unmeasured spatial correlations using spatial econometric methods among the neighboring spatial units for two reasons: (a) the collection of crash data observations associated with the spatial units does not accurately reflect the nature of the underlying process that generates the sample data, which might induce measurement errors (Anselin, 2001); (b) the spatial dimensions of socio-demographic, economic or regional activities may truly represent an important aspect in model development and may help to improve the accuracy and robustness of crash prediction and avoid underestimation of standard errors for model parameters.

A key challenge associated with the consideration of spatial dependence effects in CPMs is to address the massive amounts of multi-dimensional spatial data found in crash prediction analyses. Specifically, to gain a more precise estimation of the variability in

* Corresponding author.
E-mail addresses: dongni722@foxmail.com (N. Dong), huanghelai@csu.edu.cn (H. Huang), zhengliang@csu.edu.cn (L. Zheng).

http://dx.doi.org/10.1016/j.aap.2015.05.018

parameters by considering more complex spatial proximity structures, researchers have proposed a comprehensive investigation of different spatially neighboring structures for both road-segment-level and area-wide analyses (i.e., Aguero-Valverde and Jovanis, 2010; Wang et al., 2012). As in the study by Dong et al. (2014), CPMs accounting for spatial correlation perform better than non-spatial model and also model merely considering 0–1 first order adjacency-based proximity structure. A prevalent approach employs Bayesian spatial model with conditional autoregressive prior (i.e., CAR) to address the issue of unmeasured spatial dependences. But it has been claimed to suffer from selected limitations and fail to address complex and highly nonlinear data (the curse of dimensionality) (Karlaftis and Vlahogianni, 2011; Zeng and Huang, 2014b).

Support Vector Machine (SVM), a relatively new modeling technique, is theoretically supposed to be useful and has been employed in several studies (Yu and Abdel-Aty, 2013; Li et al., 2012, 2008; Zhang and Xie, 2008). Yu and Abdel-Aty (2013) constructed SVM models to compare with Bayesian logistic regression model in real-time crash risk evaluation. The better model predictive capability associated with SVM models implies the existence of nonlinear relationship between the dependent variables and explanatory variables which could not be captured by the logistic regression models. Li et al. (2008) investigated the potential of using an SVM model to evaluate safety performance functions for motor vehicle crashes and found that SVM models provide better goodness-of-fit than negative binomial models. It was argued that SVM model has a great ability to address classification problems while producing fewer over-fitting problems and better generalization abilities. The strength of SVM probably comes from its basis on structural risk minimization, which provides a trade-off between hypothesis space complexity and the quality of fitting the training data (Vapnik, 1998). Byvatov et al. (2003) also found that SVMs are able to efficiently address a substantial number of features due to the exploitation of kernel functions, especially for high-dimension data.

Given this new line of research activity, to the best of our knowledge, little to no research has specifically worked on a fairly thorough treatment of SVM in zonal crash prediction accounting for spatial proximity effects. This motivates our interests to fill the gap by utilizing SVM model to explore the spatial proximity effects in crash prediction.

The major challenge associated with the SVM model lies in the optimal input feature subset especially in complex and highly multivariate prediction models because the choice of feature subset influences the appropriate kernel parameters and vice versa (Huang and Wang, 2006). Recent research has postulated that feature selection becomes necessary for machine-learning tasks when working with high-dimension data (Yu and Liu, 2003). Correlation-based Feature Selector (CFS) has been developed for selecting a list of candidate variables in SVMs, which may improve model fitness as well as predictive performance (Hall, 1999). Use of CFS method has greatly expanded the potential applications of the SVM model.

The objective of this study is to explore the possibility of using SVM models and CFS method for macro-level crash frequency analysis with comparatively complex spatial data structure. SVM models with radial-basis function (RBF) kernel and linear kernel are developed. Using a dataset of Hillsborough county of Florida, the model fitness and predictive performance are compared with the CAR models. Moreover, since the SVM is unable to contain a specified function to identify the effects of explanatory variables, a comprehensive sensitivity analysis is carried out to capture the impacts of explanatory variables on the mean predicted probabilities of crash occurrence.

2. Methodology

SVM model can be used to relate various zone-level risk factors to crash occurrence, while accounting for possible spatial proximity among adjacent zones. The spatial weight features are introduced to reflect the overall spatial proximity relationships of the traffic analysis zones (TAZs), which are considered as input vectors into a SVM model in improving the predictive accuracy in this study. For comparison purposes, we also develop CAR model based on the same dataset. They are briefly described in this section, followed by the presentation of the goodness of fit measures for model comparison.

3. SVM model

For this study, the ν-SVM is employed, which has been proposed by Schölkopf et al. (2000). Specifically, the data is separated into a training set and a testing set. The ν-SVM model produces a learning model based on the training set and subsequently makes predictions on the testing set. The ν-SVM model learns the relations between the TAZs-level crash frequency and explanatory variables based on the training dataset.

Assume the training input is defined as vectors $\mathbf{x}(i) \in R^{\text{In}}$ for $i = 1, \ldots, N$, which represents the full set of zone-level contributing factors of each TAZ including road and traffic characteristics, trip production/attraction, and demographic and socioeconomic, and the training output is defined as $\mathbf{y}(i) \in R^1$ for $i = 1, \ldots, N$, which represents the crash frequency that occurred in the TAZ. The ν-SVM maps $\mathbf{x}(i)$ into a feature space $R^h > \text{In}$ with higher dimension using a function $\Phi(\mathbf{x}(i))$ to linearize the nonlinear relation between $\mathbf{x}(i)$ and $\mathbf{y}(i)$. The estimation function of $\mathbf{y}(i)$ is

$$\hat{y} = f(\mathbf{x}) = w^T\Phi(\mathbf{x}) + b$$

where $w \in R^h$ and $b \in R^1$ are coefficients. Schölkopf et al. (2000) showed that the coefficients can be determined by solving the following optimization problem:

$$\text{MinZ}(w, \varepsilon, \xi_i, \xi_i^*) = \frac{1}{2}w^T w + C\{\nu\varepsilon + \frac{1}{N}\sum_{i=1}^{N}(\xi_i + \xi_i^*)\}$$

subject to

$$w^T\Phi(\mathbf{x}(i)) + b - \mathbf{y}(i) \le \varepsilon + \xi_i \forall i = 1, \ldots, N$$

$$\mathbf{y}(i) - w^T\Phi(\mathbf{x}(i)) - b \le \varepsilon + \xi_i^* \forall i = 1, \ldots, N$$

$$\xi_i, \xi_i^* \ge 0 \quad \forall i = 1, \ldots, N$$

$$\varepsilon \ge 0$$

where ξ_i, ξ_i^* are slack variables, C is a regularization parameter, and ν is a second parameter. For each $\mathbf{x}(i)$ the allowable error is ε. Slack variables ξ_i, ξ_i^* capture the errors above ε and are penalized in the objective function via a regularization constant C.

Therefore, the estimated function of $\mathbf{y}(i)$ becomes

$$\begin{aligned}\hat{\mathbf{y}} = f(\mathbf{x}) &= \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)\Phi(\mathbf{x}(i))^T\Phi(\mathbf{x}) + b \\ &= \sum_{i=1}^{N}(\alpha_i^* - \alpha_i) \times K(\mathbf{x}(i), \mathbf{x}(j)) + b\end{aligned}$$

where $K(\mathbf{x}(i), \mathbf{x}(j)) = \Phi(\mathbf{x}(i))^T\Phi(\mathbf{x}(j))$ is the kernel function, α_i and α_i^* ares. In this study, the RBF kernel and linear kernel were considered: