Transportation Research Part C 26 (2013) 160–169

Contents lists available at SciVerse ScienceDirect

Transportation Research Part C

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

On feature selection for traffic congestion prediction

Su Yang [*]

College of Computer Science and Technology, Fudan University, Shanghai 201203, China

ARTICLE INFO

Article history:
Received 2 May 2012
Received in revised form 7 June 2012
Accepted 28 August 2012

Keywords:
Traffic congestion prediction
Pattern classification
Feature selection
Feature ranking

ABSTRACT

Traffic congestion prediction plays an important role in route guidance and traffic management. We formulate it as a binary classification problem. Through extensive experiments with real-world data, we found that a large number of sensors, usually over 100, are relevant to the prediction task at one sensor, which means wide area correlation and high dimensionality of the data. This paper investigates the first time into the feature selection problem for traffic congestion prediction. By applying feature selection, the data dimensionality can be reduced remarkably while the performance remains the same. Besides, a new traffic jam probability scoring method is proposed to solve the high-dimensional computation into many one-dimensional probabilities and its combination.

1. Introduction

Traffic congestion prediction plays an important role in intelligent transportation. For instance, the GPS navigation products equipped with traffic congestion prediction module can make more practical routing decision. Besides, the ability of traffic congestion prediction allows the traffic management department to do better management. There are three prediction problems regarding traffic congestions: Traffic volume prediction, traffic congestion prediction, and travel time prediction. This study is focused on traffic congestion prediction, which is formulated as a binary decision problem on whether the traffic volume will exceed a watching threshold shortly.

So far, traffic congestion prediction has been receiving much attention in the context of civil engineering as well as the information technology. Early researches are focused on single site prediction based on one-dimensional traffic time series such as the ARIMA model (Williams and Hoel, 2003) and the nearest neighbor method (Smith et al., 2002). Recently, the trend has been shifted to prediction based on spatial temporal correlations between traffic flows (Kanoh et al., 2005; Ando et al., 2006; Min and Wynter, 2011; Hu et al., 2009; Ghosh et al., 2009; Romaszko, 2010; He et al., 2010; Ben-Akiva et al., 2012), for instance, the vector ARMA model (Chandra and Al-Deek, 2009) incorporating both spatial and temporal correlations (Min and Wynter, 2011), and the spatial econometrics models focused on congestion propagation over adjacent links (Hu et al., 2009). The core of the existing methods is: They try to predict traffic congestions at one site based on the spatially and temporally correlated information from the sensors distributed on nearby roads, where the number of such sensors contributing to the prediction is referred to as data dimensionality. To the best of our knowledge, for almost all the existing methods, the data dimensionality does not exceed 100. For example, only 15 neighbors are considered in Min and Wynter (2011) and 10 sites in Ghosh et al. (2009). The reasons to limit the consideration in such a narrow field are two folds: (1) The high computational cost induced by higher dimensionality cannot be afforded by such methods. (2) Such methods are cause–effect based, that is, they trace the traffic congestions propagating along nearby roads to foresee the formation of new congestions (Hu et al., 2009). However, our experimental results contradict the widely accepted assumptions, that is, the signatures correlated to the traffic congestions at one site should exist in a very large scale, from more than 100 sensors in general.

[*] Tel./fax: +86 21 51355520.
E-mail address: suyang@fudan.edu.cn

0968-090X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2012.08.005

In another word, the information obtained from more than 100 sensors should be useful in predicting the traffic jams to appear at one site. Through the experiments with the real-world data of more than 4000 loop detectors located around the Twin Cities Metro freeways from 1 January to 22 September 2010, we found that the number of the relevant sensors to reach high-performance traffic congestion prediction at one sensor is over 100 in most cases. We explain such phenomenon as follows: According to the findings turned out from the simulation in Mazloumian et al. (2010), unbalanced vehicle distribution across the network of interest does have a remarkable correlation to traffic congestion from a global point of view at the whole network level. This accounts for why the seemingly irrelevant traffic patterns even far away can act as signals to indicate the possibility of traffic congestion occurrence at the watched site. This is straightforward in that for a specific network, if the traffic loads at considerable links are light, the other links must be undergoing heavy traffic loads due to the unbalanced traffic volume distribution. Accordingly, the global traffic patterns in a wide area can function to indicate possible congestions at a watched site. This gives rise to a new problem, that is, what is the optimal dimensionality of the input data and how to evaluate the significance of every sensor for traffic congestion prediction at a given sensor, which can be formulated as a feature selection/feature ranking problem in terms of pattern recognition. To the best of our knowledge, this is the first study investigating into the feature selection/feature ranking problem for traffic congestion prediction. In the context of pattern recognition, the goal of feature selection/feature ranking is to rank the quality of every attribute and identify the high-quality ones that contribute to improve the classification performance at most. By means of feature ranking/feature selection, irrelevant variables can be rejected and only the highly contributive features are preserved such that the classification performance can in general be improved while the data dimensionality is reduced. In terms of traffic congestion prediction, feature ranking/feature selection functions to identify the most significant features/sensors relevant to traffic jams at a given sensor so as to build a predictor with only relevant sensors data as input, which improves the prediction performance while reduces the data dimensionality.

The detailed implementation of the feature ranking/feature selection scheme is as follows: First, we identify the jam times at which the traffic volumes exceed a given threshold and the non-jam times at which the traffic volumes are much less than the threshold such that the positive and negative training data can be obtained, which are the traffic volumes prior to the jam and non-jam times with a certain time interval. Second, we make use of the *p*-test score presented in Golub et al. (1999) to rank the relevance of every sensor in the sense of predicting traffic jams at a given sensor, which is the so-called feature ranking. Third, we establish two Gaussian models from both the positive and negative samples for each sensor. Fourth, for the selected highly relevant sensors/features, we propose to score whether a jam will appear by combining the probabilities of how the input of every relevant sensor fits well into the corresponding Gaussian models. Fifth, we use a wrapper-like scheme (Kohavi and John, 1997) to select the optimal number of features for each sensor that can reach the best performance in the model selection procedure. Then, we use the additional data from 11 December 2010 to 20 September 2011 at the same city to test how the feature selection scheme performs.

Overall, the experiments confirm three points: (1) Signatures correlated to traffic jam prediction at one sensor exist in a wide area involving a large number of sensors in general. (2) Comparable or even better performance can be achieved with reduced dimensionality. (3) The optimal number of features determined by the wrapper-like scheme is not a fixed number but subject to which sensor is the target undergoing prediction, which forms more practical predictors.

The rest of this paper is organized as follows: We present the proposed methodology in Section 2. The experimental results are provided in Section 3. We conclude in Section 4.

2. The methodology

The traffic congestion prediction at a given sensor is formulated as a binary classification problem in the sense of pattern recognition. The goal is to classify the on-line state of a given sensor into two categories, namely jam or non-jam, by referring to the spatial temporal correlated signatures from a large number of sensors. The whole procedure is as follows: First, the training data are collected, where the historical data are partitioned into two sets: The jam set that contains the positive samples prior to the known traffic jams of a certain time lag, and the non-jam set consists of the negative samples prior to the known free travel times with the same time lag. Then, based on the positive and negative training samples, a predictor can be build up, which includes three modules: (1) feature ranking; (2) statistical decision; (3) determination of the optimal number of features. The functions of the three modules are as follows:

(1) For traffic jam congestion prediction at every sensor, the original input is the traffic volumes from all sensors, the dimensionality of which is very high, over thousands in general. However, the contribution of each sensor's data varies much in predicting traffic jams at the sensor of interest. The data from some sensors are more discriminative in distinguishing jam and non-jam while the data from some other sensors are not so discriminative. The goal of feature ranking is to score the discriminative power of each sensor in terms of separating positive training samples from negative ones so as to rank the "quality" of each sensor in predicting traffic jams at the sensor of interest. Once the rank of features/sensors are obtained for a given prediction task, we can make use of only the "high-quality" features in the subsequent decision making module to improve the prediction performance in terms of both precision and time complexity.

(2) Instead of applying classification directly, in this study, we only present the probability regarding whether a jam will occur after a certain time, which results from a statistics based method. In the learning phase, we construct two