Expert Systems With Applications 82 (2017) 301–316

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Topology-regularized universal vector autoregression for traffic forecasting in large urban areas

Florin Schimbinschi[a,*], Luis Moreira-Matias[b], Vinh Xuan Nguyen[a], James Bailey[a]

[a] University of Melbourne, Victoria, Australia
[b] NEC Laboratories Europe, Heidelberg, Germany

ARTICLE INFO

Article history:
Received 12 October 2016
Revised 15 March 2017
Accepted 5 April 2017
Available online 7 April 2017

Keywords:
Topology regularized universal vector autoregression
Multivariate timeseries forecasting
Spatiotemporal autocorrelation
Traffic prediction
Big data
Structural risk minimization

ABSTRACT

Autonomous vehicles are soon to become ubiquitous in large urban areas, encompassing cities, suburbs and vast highway networks. In turn, this will bring new challenges to the existing traffic management expert systems. Concurrently, urban development is causing growth, thus changing the network structures. As such, a new generation of adaptive algorithms are needed, ones that learn in real-time, capture the multivariate nonlinear spatio-temporal dependencies and are easily adaptable to new data (e.g. weather or crowdsourced data) and changes in network structure, without having to retrain and/or redeploy the entire system.

We propose learning Topology-Regularized Universal Vector Autoregression (TRU-VAR) and examplify deployment with of state-of-the-art function approximators. Our expert system produces reliable forecasts in large urban areas and is best described as scalable, versatile and accurate. By introducing constraints via a topology-designed adjacency matrix (TDAM), we simultaneously reduce computational complexity while improving accuracy by capturing the non-linear spatio-temporal dependencies between timeseries. The strength of our method also resides in its redundancy through modularity and adaptability via the TDAM, which can be altered even while the system is deployed. The large-scale network-wide empirical evaluations on two qualitatively and quantitatively different datasets show that our method scales well and can be trained efficiently with low generalization error.

We also provide a broad review of the literature and illustrate the complex dependencies at intersections and discuss the issues of data broadcasted by road network sensors. The lowest prediction error was observed for TRU-VAR, which outperforms ARIMA in all cases and the equivalent univariate predictors in almost all cases for both datasets. We conclude that forecasting accuracy is heavily influenced by the TDAM, which should be tailored specifically for each dataset and network type. Further improvements are possible based on including additional data in the model, such as readings from different metrics.

1. Introduction

Expert systems are at the forefront of intelligent computing and 'soft Artificial Intelligence (soft AI)'. Typically, they are seamlessly integrated in complete business solutions, making them part of the core value. In the current work we propose a system for large-area traffic forecasting, in the context of the challenges imposed by rapidly growing urban mobility networks, which we outline in the following paragraphs. Our solution relies on the formulation of a powerful inference system which is combined with expert domain knowledge of the network topology, and that can be seamlessly integrated with a control schema.

Fully autonomous traffic implies an *omniscient* AI which is comprised of two expert systems, since it has to be able to both perceive and efficiently control traffic in real time. This implies the observation of both the network state and the entities on the network. Therefore, sensing (perception) can be done via (i) passive sensors (e.g. induction loops, traffic cameras, radar) or (ii) mobile ones (e.g. Global Positioning Systems (GPS), Bluetooth, Radio Frequency Identification (RFID)). While the crowdsourced data from moving sensors (ii) can provide high-granularity data to fill accurate Origin-Destination (O-D) matrices, their penetration rate is still scarce to scale up (Moreira-Matias, Gama, Ferreira, Mendes-Moreira, & Damas, 2016).

* Corresponding author.
E-mail addresses: florin.schimbinschi@unimelb.edu.au, florinsch@student.unimelb.edu.au (F. Schimbinschi), luis.matias@neclab.eu (L. Moreira-Matias), vinh.nguyen@unimelb.edu.au (V.X. Nguyen), baileyj@unimelb.edu.au (J. Bailey).

http://dx.doi.org/10.1016/j.eswa.2017.04.015
0957-4174/

Forecasting traffic is a function of control as well, since changing traffic rules or providing route recommendations can have an impact on the network load. However, there are factors that are not a function of control, such as human error or extreme weather conditions, which are the actual unforeseen causes of congestion. Therefore, during the transition to fully autonomous traffic control, there will be an even greater need for accurate predictions. There are also many possible intelligent applications such as a personalized copilots making real time route suggestions based on users preferences and traffic conditions, economical parking metering, agile car pooling services, all of these paving the way towards fully autonomous self driving cars. Not surprisingly, the work in simulation by Au, Zhang, and Stone (2015) has shown that semi-autonomous intersection management can greatly decrease traffic delay in mixed traffic conditions (no autonomy, regular or adaptive cruise control, or full autonomy). This is possible by linking cars in a semi-autonomous way, thus solving the congestion 'wave' problem, if most of the vehicles are semi-autonomous.

Traffic prediction will therefore become paramount as urban population is growing and autonomous vehicles will become ubiquitous for both personal and public transport as well as for industrial automation. Currently, one may argue that automatic traffic might be a self-defeating process. A common scenario might be in the case when the recommendations from a prediction expert system are identical for all users in the network. In this case, new congestions can and will be created (most vehicles take the same route), which in turn invalidate the forecasts. This is evidently caused by *poor control policies* or a lack of adequate infrastructure. Fortunately, simple solutions for both of these issues exist, here we refer the reader to two references for each potential issue. Çolak, Lima, and González (2016) formulate the control problem as a collective travel time savings optimization problem, under a centralized routing scheme. Different quantified levels of social good (vs. greedy individual) are tweaked in order to achieve significant collective benefits. A simple (but more socially challenging) way to overcome the infrastructure problem is recommendations for car pooling as suggested by Guidotti, Nanni, Rinzivillo, Pedreschi, and Giannotti (2016).

Concerning the traffic prediction literature, most research effort is focused on motorways and freeways (Ahn, Ko, & Kim, 2015; Asif et al., 2014; Hong et al., 2015; Ko, Ahn, & Kim, 2016; Lippi, Bertini, & Frasconi, 2013; Lv, Tang, & Zhao, 2009; Stathopoulos & Karlaftis, 2003; Su, Dong, Jia, Qin, & Tian, 2016; Wang, Papageorgiou, & Messmer, 2008; Wu, Ho, & Lee, 2004; Zheng, Lee, & Shi, 2006), while other methods are only evaluated on certain weekdays and / or at particular times of the day (Su et al., 2016; Wu, Chen, Lu, & Yang, 2016). These methods usually deploy univariate statistical models that do not take into consideration all the properties that can lead to satisfactory generalization accuracy in the context of growth and automation in urban areas, namely: (1) real-time (online) learning; (2) model nonlinearity in the spatio-temporal domain; (3) low computation complexity and scalability to large networks; (4) contextual spatio-temporal multivariable regression via topological constraints; (5) versatility towards a broad set of infrastructure types (urban, suburban, freeways); (6) adaptation to changes in network structure, without full-network redeployment; (7) redundancy and customization for each series and adjacency matrix; (8) encoding time or using multi-metric data.

In the current work we address these issues and propose a multivariate traffic forecasting method that can capture spatio-temporal correlations, is redundant (fault tolerant) through modularity, adaptable (trivial to redeploy) to changing topologies of the network via its modular topology-designed adjacency matrix (TDAM). Our method can be efficiently deployed over large networks of broad road type variety with low prediction error and therefore generalizes well across scopes and applications. We also show (Fig. 12) that our method can predict within reasonable accuracy even up to two hours in the future – the error increases linearly and the increase rate depends on the function approximator, the TDAM and the quality of the data. We provide a comparison with state of the art methods in Table 1 according to properties that we believe are essential to the next generation of intelligent expert systems for traffic forecasting:

Our contributions are as follows: (i) We propose learning Topology-Regularized Universal Vector Autoregression (TRU-VAR), a novel method that can absorb spatio-temporal dependences between multiple sensor stations; (ii) The extension of TRU-VAR to nonlinear universal function approximators over the existing state of the art machine learning algorithms, resulting in an exhaustive comparison; (iii) Evaluations performed on two large scale real world datasets, one of which is novel; (iv) Comprehensive coverage of the literature, and an exploratory analysis considering data quality, preprocessing and possible heuristics for choosing the topology-designed adjacency matrix (TDAM).

Our conclusions are: TRU-VAR shows promising results, scales well and is easily deployable with new sensor installations; careful choice of the adjacency matrix is necessary according to the type of dataset used; high resolution data (temporal as well as spatial) is essential; missing data should be marked in order to distinguish it from real congestion events; given that the methods show quite different results on the two datasets we argue that a public set of large-scale benchmark datasets should be made available for testing the prediction performance of novel methods.

2. Related work

Traffic forecasting methodologies can be challenging to characterize and compare due to the lack of a common set of benchmarks. Despite the numerous methods that have been developed, there is yet none that is modular, design-flexible and adaptable to growing networks and changing scopes. The scope (e.g. freeway, arterial or city) and application can differ across methods. Therefore, it is not trivial to assess the overall performance of different approaches when the datasets and metrics differ. Often, subsets of the network are used for evaluating performance as opposed to the general case of network-wide prediction, which includes highways as well as suburban and urban regions. Furthermore, off-peak times and weekends are also sometimes excluded. For critical reviews of the literature we point the reader to Oh, Byon, Jang, and Yeo (2015), Vlahogianni, Karlaftis, and Golias (2014), Van Lint and Van Hinsbergen (2012), Vlahogianni, Golias, and Karlaftis (2004), Smith, Williams, and Oswald (2002) and Smith and Demetsky (1997).

Traffic metric types for sensor loops and floating car data: When it comes to metrics, speed, density and flow can be used as target prediction metrics. Flow (or volume) is the number of vehicles passing through a sensor per time unit (usually aggregated in 1, 5 or 15 min intervals). Density is the number of vehicles per kilometre. It was shown (Clark, 2003) that multi-metric predictors can result in lower prediction error. That is, variety of input data metrics is beneficial. As to the metric being predicted, some authors argue that flow is more important due to its stability (Levin & Tsao, 1980) while others (Dougherty & Cobbett, 1997) have found that traffic conditions are best described using flow and density as opposed to speed, as output metric. Nevertheless, there is a large amount of work where speed is predicted, as opposed to flow or density (Asif et al., 2014; Dougherty & Cobbett, 1997; Fusco, Colombaroni, Comelli, & Isaenko, 2015; Kamarianakis, Shen, & Wynter, 2012; Lee, Kim, & Yoon, 2007; Mitrovic, Asif, Dauwels, & Jaillet, 2015; Park et al., 2011; Salamanis, Kehagias, Filelis-Papadopoulos, Tzovaras, & Gravvanis, 2016). This data can come from either loop sensors (two are needed) or floating