Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 494 – 503

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Lattice-based spatio-temporal ensemble prediction

Saulius Samulevičius[a,*], Yoann Pitarch[b], Torben Bach Pedersen[a]

[a]*Department of Computer Science, Aalborg University, Selma Lagerlöfs Vej 300, 9220, Denmark*
[b]*Université of Toulouse, CNRS, IRIT UMR5505, F-31071, France*

Abstract

With the rapidly increasing deployment of sensor networks, large amounts of time series data are generated. One of the main challenges when dealing with such data is performing accurate predictions in order to address a broad class of application problems, ranging from mobile broadband network (MBN) optimization to preventive maintenance. To this end, time series prediction has been widely addressed by the statistics community. Nevertheless, such approaches fail in performing well when the data are more context-dependent than history-dependent. In this paper, we investigate how latent attributes can be built upon the time series in order to define a spatio-temporal context for predictions. Moreover, such attributes are often hierarchical, leading to multiple potential contexts at different levels of granularity for performing a given prediction. In support of this scenario, we propose the Lattice-Based Spatio-Temporal Ensemble Prediction (LBSTEP) approach, which allows modeling the problem as a multidimensional spatio-temporal prediction. Given an ensemble prediction model, we propose a solution for determining the most appropriate spatio-temporal context that maximizes the global prediction metrics of a set of the time series. LBSTEP is evaluated with a real-world MBN dataset, which exemplifies the intended general application domain of time series data with a strong spatio-temporal component. The experimental results shows that the proposed contextual and multi-granular view of the prediction problem is effective, in terms of both several optimization metrics and the model calculation.

Peer-review under responsibility of KES International.

Keywords: Time series; prediction; spatio-temporal; ensemble; context; lattice; hierarchy; latent attributes

1. Introduction

Time series are one of the most prominent types of data nowadays. The massive increase of sensor network deployments, e.g., in smart cities, means that a tremendous number of time series are generated. To fully benefit from this potentially highly valuable data, one of the main challenges when dealing with the time series is performing an accurate estimation of the future values. Indeed, predicting time series allows addressing a broad class of application problems ranging from mobile broadband network (MBN) optimization to preventive maintenance.

The statistics community has addressed the time series prediction problem, a.k.a. forecasting, for decades. Multiple prediction strategies have been developed, ranging from well-known state of the art techniques, *e.g.,* AutoRegressive

* Corresponding author. Tel.: +0045-9940-3574 ; fax: +0045-9940-9798.
E-mail address: sauliuss@cs.aau.dk

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.130

Integrated Moving Average (ARIMA), which can deal with a broad range of prediction problems, to specific models, *e.g.,* EGRV[1], designed for accurate energy demand forecasts. Nevertheless, such approaches fail in performing well when data are more dependent on *context* than history, especially when the context is spatio-temporal. For instance, let us consider hourly aggregated traffic in an MBN. In this scenario, the context is the most important for performing accurate predictiona. For instance, considering the type of the day, *e.g.,* weekday or week-end, the hour of the day, as well as the node location might significantly impact the traffic load prediction, *i.e.,* traffic will very likely be low in a shopping area during the night. The role of latent attributes, *i.e.,* attributes that are built upon a given dataset, is most often decisive in the success of data mining or machine learning techniques[2]. From now we will call a *context* of a time series value a set of latent attributes values that spatially, temporally, etc., characterize this value.

In a previous work[3], we have proposed STEP (Spatio-Temporal Ensemble Prediction), that starts considering this contextual aspect for unidimensional time series predictions. Roughly speaking, given a set of time series, each was considered separately and models were built for each hour and network node set. The proposed models were based on an ensemble strategy and are further extended in this paper. Nevertheless, attributes in a multidimensional context are often hierarchical. For instance, timestamps can be aggregated to hours that can further aggregated to either morning, afternoon, or night. Thus, considering that all attributes belonging to the context can be hierarchical, the main challenge is to determine which *combination of levels* is the best for achieving the most accurate prediction. Typically, STEP does not address this issue and forces the user to determine a priori this combination of levels of granularity.

In this paper, we go one step further by proposing LBSTEP (Lattice-Based Spatio-Temporal Ensemble Prediction), a multidimensional and multi-granular model for contextual attributes and investigate how to select the most appropriate models to perform the prediction. LBSTEP extends the STEP approach by (1) shifting from a single pre-defined context defined for unidimensional time series to varying contexts defined for uni/multidimensional time series, (2) proposing new ensemble strategies for combining the separate predictions, (3) providing heuristics for selecting the optimal model, and (4) providing new quality measures to accordingly select the most appropriate contexts. LBSTEP is targeted at time series with a strong spatio-temporal component. It has been validated on a real MBN dataset and the results show the effectiveness of LBSTEP, both quantitatively and qualitatively.

Section 2 presents a running example that is used throughout the rest of the paper. Section 3 introduces the definitions while Section 4 presents the models for the multidimensional contexts. Our LBSTEP approach is developed in Section 5 and validated in Section 6. Finally, related work is discussed in Section 7 and some conclusions and perspectives are drawn in Section 8.

2. Running example

A Mobile Broadband Network (MBN) is composed of nodes (cells), each providing coverage for a limited area. Constant MBN development requires additional node deployment which creates overlapping areas and allows network optimization[4]. MBN traffic varies a lot, reaching the maximum network load levels only for a limited time. MBN operators monitor many network parameters, *e.g.,* the number of active users, traffic served by the node, etc. Collected MBN data can be used for network optimization, *i.e.,* some network nodes can potentially be turned off during low load periods. Due to operational costs, it is *infeasible* to optimize MBNs based on only the *current traffic* level. Instead, MBN load *prediction* must be considered to achieve good network optimization[3]. Figure 1a provides hourly traffic measurements in MBs for a single node for 6 consecutive days. During the hours when the node carries less than 30 MB, the node can potentially be turned off. In Figure 1b, 24 consecutive Saturday traffic measurements of the selected node are provided. The dashed line shows traffic load changes. The straight line at 30MB splits traffic into two node load levels, *i.e.,* unfilled triangles present when the traffic load is *high* and the node should remain *turned on*, and filled green triangles indicate *low* traffic periods when the node *potentially could be turned off*. We notice that MBN energy potentially could be saved between 2AM and 9AM on Saturday. Considering this observation, some questions naturally arise: (1) "*Is this behavior observable for this network node only?*", (2) "*Is the behavior observable every Saturday, on weekdays, or week-end days?*", and more generally (3) "*Can a spatio-temporal context be extracted such that records belonging to this context will share the same behavior and benefit from a single prediction model?*". This paper provides a way to efficiently answer such questions.