Neurocomputing 156 (2015) 268–279

Contents lists available at ScienceDirect

Neurocomputing

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

Training subset selection in Hourly Ontario Energy Price forecasting using time series clustering-based stratification

Karol Lina López [a], Christian Gagné [a,*], Germán Castellanos-Dominguez [b], Mauricio Orozco-Alzate [b]

[a] Computer Vision and Systems Laboratory (CVSL), Département de génie électrique et génie informatique, Université Laval, Québec (Québec), Canada G1V 0A6

[b] Control and Digital Signal Processing Group, Department of Electrical and Computer Engineering, Universidad Nacional de Colombia, Manizales (Caldas), Colombia

ARTICLE INFO

Article history:
Received 29 October 2013
Received in revised form
10 November 2014
Accepted 26 December 2014
Communicated by Wei Chiang Hong
Available online 8 January 2015

Keywords:
Stratification
Data selection
Stratified sampling
Forecasting models
Hourly Ontario Energy Price

ABSTRACT

Training a given learning-based forecasting method to a satisfactory level of performance often requires a large dataset. Indeed, any data-driven methods require having examples that are providing a satisfactory representation of what we wish to model to work properly. This often implies using large datasets to be sure that the phenomenon of interest is properly sampled. However, learning from time series composed of too many samples can also be a problem, given that the computational requirements of the learning algorithms can easily grow following a polynomial complexity according to the training set size. In order to identify representative examples of a dataset, we are proposing a methodology using clustering-based stratification of time series to select a training data subset. The principle for constructing a representative sample set using this method consists in selecting heterogeneous instances picked from all the various clusters composing the dataset. Results obtained show that with a small number of training examples, obtained through the proposed clustering-based stratification, we can preserve the performance and improve the stability of models such as artificial neural networks and support vector regression, while training at a much lower computational cost. We illustrate the methodology through forecasting the one-step ahead Hourly Ontario Energy Price (HOEP).

1. Introduction

As time series are generally produced through regular sampling of a given phenomenon over a period of time, it is common to obtain very large set of redundant data using a relatively high sampling frequency (e.g., a sample every minute) over a long period of time (e.g., several years). A large training set increases the memory and processing required to generate the forecasting function. This problem can be particularly acute in situations requiring the repeated generations of a forecasting function from the data set (e.g., adjusting the hyper-parameters to learn a given forecasting function). There is thus considerable interest in reducing the training set size to remove redundancy in a training set, which can improve the space and time efficiency of the forecast models.

To reduce the training set, we propose a stratified sampling of an input space time series by clustering of the data based on a state representation of each instance. Particularly, we investigate the problem of selecting a subset of available candidate examples so as to obtain a representative description of a large dataset, in order to conduct supervised learning. We aim at removing redundancy in the training set by assuming that with a representative subset of examples, we can obtain a generalisation error close to the one obtained with the full data set. For that purpose, we assume that we already have a sufficient amount of representative data instances. We make a deterministic selection of representative examples from the different clusters to be used further in forecasting model training. This clustering-based stratification is concretely carried out for Hourly Ontario Energy Price (HOEP) forecasting. Lagged values of the HOEP as well as the lagged values of the Hourly Ontario Demand (HOD) are considered as explanatory variables.

The paper is organised as follows: Section 2 is an overview of relevant work concerning the proposed approach. In Section 3, we describe the proposed methodology of clustering-based stratification so as to select the training subsets. In Section 4, the Hourly Ontario Energy Price data set is presented, as well as the experimental set-up

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2014.12.052
0925-2312/

to evaluate performance of artificial neural networks and support vector regression trained on the data subsets generated by clustering-based stratification. Results and discussions are presented in Section 5, followed by some conclusions in Section 6.

2. Related work

The major source of inspiration of our own work originates from [1], where two methods for constructing the cross-validation folds from a dataset are presented to deterministically assess classifiers. The folds are constructed using unsupervised stratification by exploiting the instance distribution in the input space. The first proposed approach ranks the samples according to their distance to the data set centroid, and then this distance it used for partitioning. The second approach clusters and sorts the data (using the well-known *K*-means algorithm [2]) according to their cluster centre in order to conduct the partitioning. Since both methods attempt to construct more representative allocation of observations into folds, they reduce the bias of the resulting estimator.

Nonetheless, the scope of the current work is to extract the representative data subset for regression-type analysis in a context of time series forecasting, for which the explanatory variable depends on its own history. The starting point is to embed the data into a time-delayed space of suitable dimension [3]. Specifically, time series data are represented by a data point sequence typically measured at successive moments sampled over uniform time intervals. In that case, it is common to collect a large number of redundant observations. Consequently, it is important to choose a small but representative subset of training examples in order to reduce the computational burden while preserving performances and possibly improving stability.

Active learning [4] is also closely related to the current work, although dataset selection is made on-the-fly during the training. The idea of active learning is to query, and eventually label, the data samples dynamically during the learning phase. The selection of the next training samples is carried out according to some criterion, for example the level of uncertainty the learner has on the data available in the pool. Active learning is generally considered useful when all data are not labelled and the labelling operation has a given cost, as the method is able to limit greatly the number of samples requiring labelling.

Another method based on an Artificial Neural Network (ANN) for selecting examples was proposed by [5]. In particular, patterns are grouped into pairs located on both sides of a classification boundary by considering the Hamming distance. To improve the ANN generalisation ability, training is accomplished as suggested in [6]. Namely, the network training is initiated with a small subset. During the training process, generalisation of the network is estimated using an independent test set and a new pattern is selected when the generalisation estimate exceeds the apparent network error on the current training set. The new training example is selected to have the maximal error. A similar algorithm was developed by [7], called *active selection of training sets*. New patterns having the maximal error are added to the current subset using an integrated mean square error estimate. The main focus lies on the reduction of the training set size exploiting information obtained from the model due to learning from previous examples. Likewise, Leisch et al. [8] propose *cross validation with active pattern selection* based on leave-one-out cross-validation of ANN.

On the other hand, Wang et al. [9] introduce two new data selection methods to train Support Vector Machines (SVMs) for classification: the first one selects training data based on an introduced statistical confidence measure, whereas the second one uses the Hausdorff distance measure as a criterion to decide which training examples should belong to the reduced training set. In turn, Barros et al. and Songfeng [10,11] propose a procedure based on clustering by *K*-means to accelerate the training of SVMs. Clusters with mixed composition are likely to occur near the separation margins and they may hold some support vectors. Consequently, the number of vectors in a SVM training set is smaller and the training time can be decreased without compromising the generalisation capability.

Some dimensionality reduction methods can be used to select a subset of data if the time series is considered as a point in a *N-dimensional* space. The problem of dimensionality reduction in a time series has been addressed mainly by transform methods. Particularly, Chakrabarti et al. [12] introduced the *Adaptive Piecewise Constant Approximation* (APCA) that approximates each time series by a set of constant value segments of varying length such that their individual reconstruction errors are minimal. An et al. [13] propose an index compression method named *Grid-based Datawise Dimensionality Reduction* which attempts to preserve the characteristics of the time-series.

The method we are proposing differs from previous ones in that the selection of samples is performed directly in the input space. Moreover, we take into account the history of the time series since inputs are composed of lags. Note that, we do not consider the example selection and training of the forecast model to be conducted simultaneously, as this can be computationally expensive and depends on to the training algorithm used. Furthermore, in the procedure proposed herein, no prior knowledge of the desired outputs is required, as the method is unsupervised. Thus, it can be applied to either regression or classification problems, including those cases when labels are not available beforehand.

3. Selection of representative training data examples

The selection of representative training data examples is carried out in two steps: (1) applying the clustering procedure to the time series, in order to discover pattern behaviours on input space, and (2) selecting the data from the clusters obtained, building stratified sample sets that form a parsimonious data representation.

The main goal is to select data best representing the structure of the inputs. For that purpose, we apply clustering methods on the input space, in order to determine the different groups of similar instances. From that point on, we divert the data of a given cluster evenly into the different folds (data subsets). Clustering is achieved through the classical *K*-means algorithm along with the Euclidean distance between each instance and the associated cluster centre.

For time series forecasting, we predict the value of a given variable at the current time step using as input some of its past values predefined during time steps, termed *lagged values*. The lagged values of the HOEP as well as the lagged values of the HOD are considered as explanatory variables [14]. More formally, let the list $\boldsymbol{a} \in \mathbb{R}^{1\times n}$ be the n lags of the HOEP and $\boldsymbol{b} \in \mathbb{R}^{1\times m}$ be the m lags of the HOD at current time t used to build up the list $\boldsymbol{l}(t) \in \mathbb{R}^{1\times n+m}$, which represents the lagged values (*Input Space*) of the forecasted variable:

$$\boldsymbol{a} = [a_1, \ldots, a_n], \tag{1a}$$

$$\boldsymbol{b} = [b_1, \ldots, b_m], \tag{1b}$$

$$\boldsymbol{v}_1(t) = [\mathrm{HOEP}(t-a_1), \ldots, \mathrm{HOEP}(t-a_n)], \tag{1c}$$

$$\boldsymbol{v}_2(t) = [\mathrm{HOD}(t-b_1), \ldots, \mathrm{HOD}(t-b_m)], \tag{1d}$$

where $\boldsymbol{v}_1(t)$ are the HOEP lagged values and $\boldsymbol{v}_2(t)$ are the HOD lagged values.