Neurocomputing 141 (2014) 236–245

Contents lists available at ScienceDirect

Neurocomputing

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

A methodology for training set instance selection using mutual information in time series prediction

Miloš B. Stojanović [a,*,1,2,3], Miloš M. Božić [b], Milena M. Stanković [b], Zoran P. Stajić [b]

[a] College of Applied Technical Sciences, Aleksandra Medvedeva 20, 18000 Niš, Serbia
[b] Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia

ARTICLE INFO

Article history:
Received 21 September 2012
Received in revised form
25 November 2013
Accepted 19 March 2014
Communicated by: P. Zhang
Available online 8 April 2014

Keywords:
Instance selection
Mutual information
Time-series prediction

ABSTRACT

Training set instance selection is an important preprocessing step in many machine learning problems, including time series prediction, and has to be considered in practice in order to increase the quality of the predictions and possibly reduce training time. Recently, the usage of mutual information (MI) has been proposed in regression tasks, mostly for feature selection and for identifying the real data from data sets that contain noise and outliers. This paper proposes a new methodology for training set instance selection for long-term time series prediction. The proposed methodology combines a recursive prediction strategy and advanced instance selection criterion—the nearest neighbor based MI estimator. An application of the concept of MI is presented for the selection of training instances based on MI computation between initial training set instances and the current forecasting instance, for every prediction step. The novelty of the approach lies in the fact that it fits the initial training subset with the current forecasting instance, and consequently reduces the uncertainty of the prediction. In this way, by selecting instances which share a large amount of MI with the current forecasting instance in every prediction step, error propagation and accumulation can be reduced, both of which are well known shortcomings of the recursive prediction strategy, thus leading to better forecasting quality. Another element which sets this approach apart from others is that it is not proposed as an outlier detector, but for the instance selection of data which do not necessarily have to contain noise and outliers. The results obtained from the data sets from NN5 competition in time series prediction indicate that the proposed method increases the quality of long-term time series prediction, as well as reduces the amount of instances needed for building the model.

1. Introduction

Time series forecasting is nowadays a key topic in various fields. In financial economics, stock exchange courses are predicted [1], in computer science, the flow of data through networks and frequency of access to websites [2], in power systems, distribution companies forecast the load of the following day [3], etc. Conventional methods for time series forecasting, developed during the 1970s and 80s, among which the most popular were the linear regression [4], Box-Jenkins ARIMA [5] and exponential smoothing [6], cannot always provide an accurate and unbiased estimation of time series in cases when the underlying system processes are typically non-linear, non-stationary and not defined a priori. In addition, very often the choice of the prediction method and the determination of its parameters depend on knowing the properties of the underlying process. In order to address these problems, machine learning models, among which the most popular were artificial neural networks (ANNs) [7] and support vector machines (SVMs) [8], have established themselves in the last two decades as serious contenders to classic statistical models in the area of forecasting. Basically, time series prediction in machine learning comes down to training the model which establishes a mapping between training set instances and their target values. Then, the trained model estimates future values based on current and past data samples. The determination of sufficient and necessary data is essential for training a good forecasting model. If the amount of data in the training set is

* Corresponding author. Tel.: +381 018 4233099 (Home), +381 065 8136718 (Mobile).
E-mail addresses: milosstojanovic10380@yahoo.com, milos.stojanovic@vtsnis.edu.rs (M.B. Stojanović).
URL: http://www.vtsnis.edu.rs/index_english.html (M.B. Stojanović).
[1] Home address: Bulevar doktora Zorana Đinđića 29/5, 18000 Niš, Serbia.
[2] Job address: Aleksandra Medvedeva 20, 18000 Niš, Serbia.
[3] Institution (job): College of Applied Technical Sciences, Niš, Serbia.

http://dx.doi.org/10.1016/j.neucom.2014.03.006
0925-2312/

insufficient, the forecasting of the model will be poor, and the model may be prone to underfitting. On the other hand, if the training set is too large, the information that it provides to the model could be unnecessary or redundant. As a result, the model could have poor generalization performance and may be prone to overfitting. Based on the size of the prediction horizon of a problem, predictions can be classified into two categories: short-term and a long-term. Take for example electric load forecasting. If hourly predictions for one day ahead are needed, it is a short-term forecasting problem requiring values for the next 24 steps. On the other hand, if yearly predictions for the next 10 or 20 years are needed, it is a long-term forecasting problem, requiring predictions for the next 10 or 20 steps. But, both of them are classified as long-term time series prediction, based on the number of steps in the prediction horizon needed to be predicted. When one-step ahead prediction is needed it is referred to short-term prediction. But when multi-step ahead predictions are needed, it is called a long-term time-series prediction. Long-term predictions are especially challenging, where multiple steps ahead have to be predicted. The problem that occurs in long-term prediction is that uncertainty increases with an increase in the number of steps in the prediction horizon. It depends on several factors, such as the accumulation of errors and lack of information. Moreover, with the increase in the number of steps that need to be predicted, a model selection problem emerges, because the environment in which the model was developed may change over time [9]. Another problem that affects the quality of long-term predictions occurs when the time series consist of daily or shorter time intervals, i.e. if they contain high frequency data [10]. High frequency data represents a specific type of forecasting problem, rendering conventional methods inappropriate and demanding new approaches [11].

The selection of an appropriate subset of instances that are included in the initial training set is a very important preprocessing step, especially in long-term time series prediction tasks. It may provide improvements in terms of the quality of the output results and in the reduction of computational time. This problem in literature is considered from several different perspectives. Instance selection can be approached from the aspect of removing outliers and noise from distorted data sets, as presented in [12–14]. Instance selection of data that contain outliers aims to remove elements from the training set that in some way differ from most other elements in the input set. From another perspective, this problem can be considered 'data shifting', where the joint distribution of inputs and outputs differs between the training and test stage, as presented in [15–17]. It usually appears in non-stationary environments, when the training environment is different from the test one, whether due to a temporal or a spatial change. There are also various methods based on active learning that deal with the selection of relevant instances, some of which can be found in [18–20]. The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer training instances if it is allowed to choose the data from which it learns. Active learning is also closely related to covariate shift, where the training input distribution is different from the test input distribution [21]. Finally, the instance selection of data which do not necessarily have to contain noise and outliers, determines a subset of the initial training set. It can be used to train a more accurate model, with a possible reduction in computational time.

In order to perform the selection of instances that the learning algorithm will use, three main approaches have been used: the incremental, decremental and batch, as presented in [22–24]. While in the incremental approach the selection algorithm starts from an empty set of instances, and adds them iteratively, in the decrement approach selection the algorithm starts from a full set of available instances, and removes those which did not meet the predefined selection criterion. The batch method performs several iterations through the initial training set before removing one instance. In each iteration it marks instances that are candidates to be removed in the next iteration. Recently, the application of evolutionary algorithms, boosting techniques and pruning techniques have been used to tackle this problem [25–27]. According to the selection strategy, instance selection can be tackled with filter and wrapper methods, as presented in [28,29]. In the filter methods, the selection criterion uses a selection function which is independent from the training algorithm used to form the regression model. On the other hand, in the wrapper methods the selection criterion is based on the evaluation measure obtained by the regression model. In other words, it is embedded into the evaluation function of the model. In this way, instances that do not contribute to the prediction quality are discarded from the training set.

Most research on instance selection, which has been done so far, refers to classification problems [30], while only a few papers deal with instance selection in regression tasks, especially in the case of long-term time series prediction. For example, [31] shows a method of k-surrounding neighbors for the selection of input vectors, while the output is calculated with the k-nearest neighbors (kNN) algorithm. In [13] a genetic algorithm is presented to perform feature and instance selection for linear regression models. In [32,33] a new distance function, which integrates the Euclidean distance and the dissimilarity of the trend of a time series, is defined as a similarity measure between two instances for long-term time series prediction. By selecting similar instances in the training set for each testing instance based on the modified kNN approach, prediction quality can be increased. Only recently a mutual information (MI) estimator based on nearest neighbors, which allows MI estimation directly from the data set, has been introduced for instance selection in time series prediction. Its aim is to remove outliers and noise from highly distorted data sets [14,34]. The applied algorithm determines the loss of MI with respect to its neighbors in such a way that if a loss of MI is similar to the instance near the studied one, then this instance must be included in the training dataset. This approach has proved successful in situations when it has been applied to training sets which are artificially distorted by adding noise or outliers. In [35], the concept of MI is applied for improving short-term load forecasting through the selection of instances with similar load patterns to the current forecasting instance.

The research presented in this paper is motivated by the work presented in [14,34] and represents an extension of the approach proposed in [35]. It is framed within the instance selection of data which do not necessarily have to contain noise and outliers. It proposes a methodology for training subsets selection in long-term recursive time series prediction, by using MI to decide which instances should or should not be included in the training data set. The methodology is based on a decremental approach and filter method which uses MI as the selection criterion. In this way, by selecting instances which share a large amount of MI with the current forecasting instance in every prediction step, error propagation and accumulation can be reduced. Since they are well known shortcomings of the recursive prediction strategy, this can lead to better forecasting quality. In this paper, the least squares support vector machines (LS-SVMs) were used as nonlinear models to present the application of the proposed methodology.

The rest of the paper is organized as follows: Section 2 presents the formulation of MI and describes the method used to compute it, followed by Section 3 which introduces the methodology to select training set instances. Section 4 shortly reviews the basics of LS-SVMs. Section 5 includes a variety of experiments to verify the proposed approach, and finally, Section 6 draws the conclusions.