Information Sciences 000 (2017) 1–22

Contents lists available at ScienceDirect

Information Sciences

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

Knowledge discovery in data streams with the orthogonal series-based generalized regression neural networks

Piotr Duda[a,*], Maciej Jaworski[a], Leszek Rutkowski[a,b]

[a] The Institute of Computational Intelligence, Czestochowa University of Technology, Armii Krajowej 36, Czestochowa 42-200, Poland
[b] Information Technology Institute, Academy of Social Sciences, Łódź 90-113, Poland

ARTICLE INFO

Article history:
Received 31 March 2017
Revised 6 June 2017
Accepted 11 July 2017
Available online xxx

Keywords:
Time-varying environment
Stream data
Regression function
Orthogonal series
Convergence in probability and with probability one

ABSTRACT

In this paper, a method for nonparametric regression estimation in a time-varying environment is presented. The orthogonal series-based kernels are used to design learning procedures tracking non-stationary systems changes under non-stationary noise. The presented procedures, constructed in the spirit of generalized regression neural networks, are a very effective tool to deal with stream data. The convergences in probability and with probability one are proved, experimental results are given and discussed.

1. Introduction

One of the most challenging problems in data mining is related to learning in non-stationary environments. For the recent excellent survey of these problems the reader is referred to [1]. Various methods have been developed to cope with the so-called "concept drift" in the context of designing intelligent systems, stream data mining or incremental machine learning [2–10]. Vast majority of them are devoted to pattern classification whereas only few deal with a non-stationary regression. Most of them rely on a Gaussian or Markov models, extend Support Vector Machine or Extreme Learning Machine to regression problems, implement regression trees or polynomial regression in a non-stationary environment. We will briefly describe these approaches.

A lot of work has been put to investigate the methods which treat regression as a Gaussian process. To address the problem of large-scale and non-stationary data set the authors in [11] proposed a K-Nearest-Neighbor-based Kalman filter for the Gaussian process regression (KNN-KFGP). The developed method worked in a few steps. Firstly, the test-input driven KNN mechanism, to group the training set into a number of small collections, is performed. Secondly, the latent function values of these collections are used as the unknown states and a novel state space model with the GP prior is constructed. Thirdly, the Kalman filter on this state space model, to efficiently filter out the latent function values, is explored for prediction. In a result, the KNN mechanism helps each test point to find its strongly correlated local training subset, and thus the KNN-KFGP algorithm can model non-stationarity in a flexible manner. The other consideration about the Gaussian process regression in shown in [12]. The author proposed two approaches for the on-line Gaussian process regression with low com-

* Corresponding author.
E-mail address: piotr.duda@iisi.pcz.pl (P. Duda).

http://dx.doi.org/10.1016/j.ins.2017.07.013
0020-0255/

putational and memory demands. The first approach assumes known hyper parameters and performs regression on a set of basis vectors that store mean and covariance estimates of the latent function. The second approach additionally learns the hyper parameters on-line. For this purpose, techniques from nonlinear Gaussian state estimation are exploited. More about Gaussian process regression can be found in [13–15].

The comparison of Markov switching regression, proposed in [16], and time-varying parameter methods is presented in [17]. The novelty of this paper was to select the coefficients of the detection methods by optimizing the profit objective functions of the trading activity, using statistical estimates as initial values. The paper also developed a sequential approach, based on sliding windows, to cope with the time-variability of Markov switching coefficients.

In the paper [18], a cost-efficient online adaptive learning approach is proposed for Support Vector Regression (SVR) by combining Feature Vector Selection and Incremental and Decremental Learning. In this approach, the model is adaptively modified only when different pattern drifts are detected according to proposed criteria. Two tolerance parameters are introduced in the approach to control the computational complexity, reduce the influence of the intrinsic noise in the data and avoid the overfitting problem of SVR. The same authors in [19] proposed an SVR-based ensemble model. Other approaches with applications of SVR can be found in [20–22]

Since the On-Line Sequential Extreme Learning Machine (OS-ELM) has been proposed in [23], many researchers have tried to apply this algorithm to work in a non-stationary environment. In [24], the authors developed an algorithm using the OS-ELM with an adaptive forgetting factor to improve performance in time-varying environments. A special batch variant of the ELM, extreme learning machine with kernels (ELMK), was proposed in [25]. It uses unknown kernel mappings instead of known hidden layer mappings; in consequence there is no need to select the number of hidden nodes. Another combination of the ELM and kernel methods was proposed in [26]. In [10], the batch-learning type and time-varying version of the ELM, called ELM-TV, is presented. The proposed version can deal with applications where sequential arrival or large number of training data occurs. In [27], a new sequential learning algorithm is constructed by combining the OS-ELM and Kalman filter regression.

Considerable effort has been devoted to the development of regression trees in non-stationary environments, see [28,29]. The problem of functional polynomial regression in a non-stationary environment was considered in [30,31].

In [32], the authors proposed a varying-coefficient fractionally exponential (VC-FEXP) model which allows to detect the dynamic change for both short-memory and long-memory structures. This approach is built on a semi-parametric class of models, whose specification is extended from a stationary fractionally exponential (FEXP) model by allowing parameters in the spectra to vary smoothly over time. The authors applied a time-varying version of the log-periodogram regression. Under this regression framework, they suggested a generalized goodness-of-fit test to detect various aspects of non-stationarity. Another test procedure is presented in [33]. The author proposed a Gini-based statistical test for a unit root. This test is based on the well-known Dickey Fuller test [33], where the ordinary least squares regression is replaced by the semi-parametric Gini regression in modeling the autoregressive process. The critical values are determined based on the residual-based bootstrap method. The proposed methodology takes into account variability of values and ranks. Therefore, it provides robust estimators that are rank-based, while avoiding loss of information. The Gini methodology can be used for a wide range of distributions.

It should be emphasized that listed above methods and techniques rely heavily on various heuristic approaches. Motivated by this fact, the lack of mathematically justified methods in presented above literature review, in this paper we will develop non-parametric algorithms tracking a wide spectrum of concept-drifts and possessing solid mathematical foundations.

Let $X_1, X_2 \ldots$ be a sequence of independent and identically distributed random variables in R^p with a common density function f. In this paper, we will consider two non-stationary models:

i)

$$Y_n = \rho(X_n) + Z_n, \qquad n = 1, 2, \ldots, \tag{1}$$

ii)

$$Y_n = \phi_n(X_n) + Z_n, \qquad n = 1, 2, \ldots, \tag{2}$$

where $\rho(\cdot)$, $\phi_n(\cdot)$, for $n = 1, 2, \ldots$, are unknown functions and Z_n are independent random variables with time-varying distributions such that

$$\mathbb{E}Z_n = 0, \quad \mathbb{E}Z_n^2 = d_n, \qquad n = 1, 2, \ldots. \tag{3}$$

Various examples of systems working in the presence of noise with time-varying variances can be found in [34–36].

Our problem is to design a nonparametric procedure tracking changes of unknown functions $\rho_(x)$ in model (1), and $\phi_n(x)$ in model (2), for $n = 1, 2, \ldots$, based on the observations $(X_1, Y_1), (X_2, Y_2), \ldots$. To solve the problem we propose to use a nonparametric technique based on the orthogonal series expansions of unknown functions. Our approach can be treated as a variant of generalized regression neural networks (GRNN) suggested by Specht [37] and studied by many authors in non-stream scenario, see e.g. [38,39]. Our paper differs from previous approaches in two aspects. First, contrary to the classical GRNN were all the samples must be stored, we will use recursive formulas to cope with coming stream data. Moreover, we

Please cite this article as: P. Duda et al., Knowledge discovery in data streams with the orthogonal series-based generalized regression neural networks, Information Sciences (2017), http://dx.doi.org/10.1016/j.ins.2017.07.013