Signal Processing 137 (2017) 22–32

Contents lists available at ScienceDirect

Signal Processing

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

Highly efficient hierarchical online nonlinear regression using second order methods

Burak C. Civek[a,*], Ibrahim Delibalta[b], Suleyman S. Kozat[a]

[a] Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey
[b] Turk Telekom Communications Services Inc., Istanbul, Turkey

ARTICLE INFO

Article history:
Received 29 July 2016
Revised 21 January 2017
Accepted 25 January 2017
Available online 26 January 2017

Keywords:
Hierarchical tree
Nonlinear regression
Online learning
Piecewise linear regression
Newton method

ABSTRACT

We introduce highly efficient online nonlinear regression algorithms that are suitable for real life applications. We process the data in a truly online manner such that no storage is needed, i.e., the data is discarded after being used. For nonlinear modeling we use a hierarchical piecewise linear approach based on the notion of decision trees where the space of the regressor vectors is adaptively partitioned based on the performance. As the first time in the literature, we learn both the piecewise linear partitioning of the regressor space as well as the linear models in each region using highly effective second order methods, i.e., Newton–Raphson Methods. Hence, we avoid the well known over fitting issues by using piecewise linear models, however, since both the region boundaries as well as the linear models in each region are trained using the second order methods, we achieve substantial performance compared to the state of the art. We demonstrate our gains over the well known benchmark data sets and provide performance results in an individual sequence manner guaranteed to hold without any statistical assumptions. Hence, the introduced algorithms address computational complexity issues widely encountered in real life applications while providing superior guaranteed performance in a strong deterministic sense.

1. Introduction

Recent developments in information technologies, intelligent use of mobile devices and Internet have procured an extensive amount of data for the nonlinear modeling systems [1,2]. Today, many sources of information from shares on social networks to blogs, from intelligent device activities to large scale sensor networks are easily accessible [3]. Efficient and effective processing of this data can significantly improve the performance of many signal processing and machine learning algorithms [4–6]. In accordance with the aim of achieving more efficient algorithms, hierarchical approaches have been recently proposed for nonlinear modeling systems [7,8].

In this paper, we investigate the nonlinear regression problem that is one of the most important topics in the machine learning and signal processing literatures. This problem arises in several different applications such as signal modeling [9,10], financial market [11] and trend analyses [12], intrusion detection [13] and recommendation [14]. However, traditional regression techniques show less than adequate performance in real-life applications having big data since (1) data acquired from diverse sources are too large in size to be efficiently processed or stored by conventional signal processing and machine learning methods [15–18]; (2) the performance of the conventional methods is further impaired by the highly variable properties, structure and quality of data acquired at high speeds [15–17].

In this context, to accommodate these problems, we introduce online regression algorithms that process the data in an online manner, i.e., instantly, without any storage, and then discard the data after using and learning [18,19]. Hence our methods can constantly adapt to the changing statistics or quality of the data so that they can be robust and prone to variations and uncertainties [19–21]. From a unified point of view, in such problems, we sequentially observe a real valued sequence vector sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ and produce a decision (or an action) d_t at each time t based on the past $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t$. After the desired output d_t is revealed, we suffer a loss and our goal is to minimize the accumulated (and possibly weighted) loss as much as possible while using a limited amount of information from the past.

To this end, for nonlinear regression, we use a hierarchical piecewise linear model based on the notion of decision trees, where the space of the regressor vectors, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots,$ is adaptively partitioned and continuously optimized in order to enhance the

[*] Corresponding author.
E-mail addresses: civek@ee.bilkent.edu.tr (B.C. Civek), ibrahim.delibalta@turktelekom.com.tr (I. Delibalta), kozat@ee.bilkent.edu.tr (S.S. Kozat).

http://dx.doi.org/10.1016/j.sigpro.2017.01.029
0165-1684/© 2017 Elsevier B.V. All rights reserved.

performance [10,22,23]. We note that the piecewise linear models are extensively used in the signal processing literature to mitigate the overtraining issues that arise because of using nonlinear models [10]. However their performance in real life applications are less than adequate since their successful application highly depends on the accurate selection of the piecewise regions that correctly model the underlying data [24]. Clearly, such a goal is impossible in an online setting since either the best partition is not known, i.e., the data arrives sequentially, or in real life applications the statistics of the data and the best selection of the regions change in time. To this end, as the first time in the literature, we learn both the piecewise linear partitioning of the regressor space as well as the linear models in each region using highly effective second order methods, i.e., Newton–Raphson Methods [25]. Hence, we avoid the well known over fitting issues by using piecewise linear models, moreover, since both the region boundaries as well as the linear models in each region are trained using the second order methods we achieve substantial performance compared to the state of the art [25]. We demonstrate our gains over the well known benchmark data sets extensively used in the machine learning literature. We also provide theoretical performance results in an individual sequence manner that are guaranteed to hold without any statistical assumptions [18]. In this sense, the introduced algorithms address computational complexity issues widely encountered in real life applications while providing superior guaranteed performance in a strong deterministic sense.

In adaptive signal processing literature, there exist methods which develop an approach based on weighted averaging of all possible models of a tree based partitioning instead of solely relying on a particular piecewise linear model [23,24]. These methods use the entire partitions of the regressor space and implement a full binary tree to form an online piecewise linear regressor. Such approaches are confirmed to lessen the bias variance trade off in a deterministic framework [23,24]. However, these methods do not update the corresponding partitioning of the regressor space based on the upcoming data. One such example is that the recursive dyadic partitioning, which partitions the regressor space using separation functions that are required to be parallel to the axes [26]. Moreover, these methods usually do not provide a theoretical justification for the weighting of the models, even if there exist inspirations from information theoretic deliberations [27]. For instance, there is an algorithmic concern on the definitions of both the exponentially weighted performance measure and the “universal weighting” coefficients [19,24,28,29] instead of a complete theoretical justifications (except the universal bounds). Specifically, these methods are constructed in such a way that there is a significant correlation between the weighting coefficients, algorithmic parameters and their performance, i.e., one should adjust these parameters to the specific application for successful process [24]. Besides these approaches, there exists an algorithm providing adaptive tree structure for the partitions, e.g., the Decision Adaptive Tree (DAT) [30]. The DAT produces the final estimate using the weighted average of the outcomes of all possible subtrees, which results in a computational complexity of $O(m4^d)$, where m is the data dimension and d represents the depth. However, this would affect the computational efficiency adversely for the cases involving highly nonlinear structures. In this work, we propose a different approach that avoids combining the prediction of each subtrees and offers a computational complexity of $O(m^2 2^d)$. Hence, we achieve an algorithm that is more efficient and effective for the cases involving higher nonlinearities, whereas the DAT is more feasible when the data dimension is quite high. Moreover, we illustrate in our experiments that our algorithm requires less number of data samples to capture the underlying data structure. Overall, the proposed methods are completely generic such that they are capable of incorporating all Recursive Dyadic, Random Projection (RP) and k-d trees in their framework, e.g., we initialize the partitioning process by using the RP trees and adaptively learn the complete structure of the tree based on the data progress to minimize the final error.

In Section 2, we first present the main framework for nonlinear regression and piecewise linear modeling. In Section 3, we propose three algorithms with regressor space partitioning and present guaranteed upper bounds on the performances. These algorithms adaptively learn the partitioning structure, region boundaries and region regressors to minimize the final regression error. We then demonstrate the performance of our algorithms through widely used benchmark data sets in Section 4. We then finalize our paper with concluding remarks.

2. Problem description

In this paper, all vectors are column vectors and represented by lower case boldface letters. For matrices, we use upper case boldface letters. The ℓ^2-norm of a vector $\boldsymbol{x}$ is given by $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$ where $\boldsymbol{x}^T$ denotes the ordinary transpose. The identity matrix with $n \times n$ dimension is represented by $\boldsymbol{I}_n$.

We work in an online setting, where we estimate a data sequence $y_t \in \mathbb{R}$ at time $t \geq 1$ using the corresponding observed feature vector $\boldsymbol{x}_t \in \mathbb{R}^m$ and then discard $\boldsymbol{x}_t$ without any storage. Our goal is to sequentially estimate y_t using $\boldsymbol{x}_t$ as

$$\hat{y}_t = f_t(\boldsymbol{x}_t)$$

where $f_t(\cdot)$ is a function of past observations. In this work, we use nonlinear functions to model y_t, since in most real life applications, linear regressors are inadequate to successively model the intrinsic relation between the feature vector $\boldsymbol{x}_t$ and the desired data y_t [31]. Different from linear regressors, nonlinear functions are quite powerful and usually overfit in most real life cases [32]. To this end, we choose piecewise linear functions due to their capability of approximating most nonlinear models [33]. In order to construct a piecewise linear model, we partition the space of regressor vectors into K distinct m-dimensional regions S_k^m, where $\bigcup_{k=1}^K S_k^m = \mathbb{R}^m$ and $S_i^m \cap S_j^m = \emptyset$ when $i \neq j$. In each region, we use a linear regressor, i.e., $\hat{y}_{t,i} = \boldsymbol{w}_{t,i}^T\boldsymbol{x}_t + c_{t,i}$, where $\boldsymbol{w}_{t,\,i}$ is the linear regression vector, $c_{t,\,i}$ is the offset and $\hat{y}_{t,i}$ is the estimate corresponding to the ith region. We represent $\hat{y}_{t,i}$ in a more compact form as $\hat{y}_{t,i} = \boldsymbol{w}_{t,i}^T\boldsymbol{x}_t$, by including a bias term into each weight vector $\boldsymbol{w}_{t,\,i}$ and increasing the dimension of the space by 1, where the last entry of $\boldsymbol{x}_t$ is always set to 1.

To clarify the framework, in Fig. 1, we present a one dimensional regression problem, where we generate the data sequence using the nonlinear model

$$y_t = \exp(x_t \sin(4\pi x_t)) + \nu_t,$$

where x_t is a sample function from an i.i.d. standard uniform random process and ν_t has normal distribution with zero mean and 0.1 variance. Here, we demonstrate two different cases to emphasize the difficulties in piecewise linear modeling. For the case given in the upper plot, we partition the regression space into three regions and fit linear regressors to each partition. However, this construction does not approximate the given nonlinear model well enough since the underlying partition does not match exactly to the data. In order to better model the generated data, we use the second model as shown in the lower plot, where we have eight regions particularly selected according to the distribution of the data points. As the two cases signified in Fig. 1 imply, there are two major problems when using piecewise linear models. The first one is to determine the piecewise regions properly. Randomly selecting the partitions causes inadequately approximating models as indicated in the underfitting case on the top of Fig. 1 [22]. The second problem is to find out the linear model that best fits the data in