Neurocomputing 175 (2016) 233–242

Contents lists available at ScienceDirect

Neurocomputing

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

Kernel online learning with adaptive kernel width

Haijin Fan [a,b,*], Qing Song [a], Sumit B. Shrestha [b]

[a] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore
[b] Institute of High Performance Computing, A*STAR, 1 Fusionopolis Way, #16-16 Connexis North, Singapore 138632, Singapore

ARTICLE INFO

Article history:
Received 14 April 2015
Received in revised form
27 September 2015
Accepted 20 October 2015
Communicated by Haowei Liu
Available online 3 November 2015

Keywords:
Online learning
Kernel width
Adaptive learning
Cumulative coherence
Convergence

ABSTRACT

This paper discusses a unified framework for kernel online learning (KOL) algorithm with adaptive kernels. Unlike the traditional KOL algorithms which applied a fixed kernel width in the training process, the kernel width is considered as an additional free parameter and can be adapted automatically. A robust training method is proposed based on an adaptive dead zone scheme. The kernel weight and the kernel width are updated under a unified framework, where they share the same learning parameters. We present a theoretical convergence analysis of the proposed adaptive training method which can switch off the learning when the training error is too small in terms of external disturbance. Meanwhile, in the regularization of the kernel function number, an in-depth measure concept: the cumulative coherence is applied. A dictionary with predefined size is selected by online minimization of its cumulative coherence without using any parameters related to the prior knowledge of the training samples. Simulation results show that the proposed algorithm can adapt the training data effectively with different initial kernel width. Its performance could be better in both testing accuracy and convergence speed compared with the kernel algorithms with a fixed kernel width.

1. Introduction

Kernel online learning (KOL) algorithms have been extensively studied for various applications, including time series prediction, function approximation, nonlinear regression and novelty detection, etc [1–4]. The mathematical fundamental is the use of Mercer kernels, which performs feature transformation from a low dimensional space to a high dimensional space. The high dimensionality makes many nonlinear problem linearly solvable in the Reproducing Kernel Hilbert Space (RKHS) [5]. The popular "kernel trick" makes the inner product evaluation in RKHS efficient without introducing extra heavy computational cost. The KOL algorithms are tuned in a sequential way, where at each training iteration, only one sample is employed to update the algorithms.

For the KOL algorithm, the estimated output function is a linear combination of selected kernel functions. Many linear least mean square (LMS) algorithms are extensible to the kernel algorithm in RKHS due to its linear form. The well-known kernel online learning algorithms include the kernel least mean square (KLMS) algorithm [6,7], the kernel normalized least mean square (KNLMS) algorithm and the kernel affine projection (KAP) algorithm [3]. These algorithms update the kernel weight using the gradient information and constrain the learning speed by a fixed learning rate. In the LMS typed algorithm, the learning rate plays an important role in maintaining their convergence speed and stability. In traditional LMS algorithms, usually the learning rate is small in order to make the system not divergent. To overcome the divergence problem, adaptive learning rate is required [8,9]. The KOL algorithm also suffers the same problem. A weight convergence analysis was done in the quantized kernel least mean square (QKLMS) algorithm [7,10], where the result shows that the reasonable learning rate is bounded by a small value, and by choosing a fixed learning rate, a trade-off is made between the convergence speed and testing accuracy.

In kernel online learning, the number of training samples is continuously increasing. Sparsification performs as one important operation to keep the complexity of the algorithm regularized. It helps to curb the growing kernel function number while the training sample sequentially arrives. This was usually done by checking whether the new training sample was approximately independent, novel, or informative according to different criterions. The criterions were evaluated either in the RKHS or in the original feature space, such as the approximate linear dependency (ALD) criterion [2], the coherence based criterion [3] and the surprise criterion [11], which were evaluated in the RKHS. The novelty criterion, firstly introduced in the resource-allocating network [12] by evaluating the Euclidian distance of feature vectors in the original feature space, was applied for sparsification in kernel algorithm [11]. As an alternative way of sparsification, Chen

* Corresponding author at: School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore.
E-mail addresses: hfan1@e.ntu.edu.sg, fanhiking@gmail.com (H. Fan).

http://dx.doi.org/10.1016/j.neucom.2015.10.055
0925-2312/

et al. [7] proposed a quantization approach to tune and train the kernel algorithm. The NORMA algorithm [1] applied a truncation method to keep the kernel function number constant by using the feature vector of the most recent training samples as kernel centers. The method was extended to a model based KOL algorithm by solving a constrained optimization model [13].

The kernel width also plays an important role in the performance of kernel algorithms [14,15]. A small kernel width would make the locality of data preserved while a large kernel width will make the estimation function smooth [16]. Thus at the stage of training process, an optimal kernel width can make a compromise between the locality and smoothness of the estimated function. Recently, some optimal kernel width selection methods have been developed, including the kernel target alignment [17], information theoretic method [18], kernel polarization [19,20] in classification. However, in the recently proposed KOL algorithms, a fixed kernel width was adopted. In this paper, we take the kernel width into consideration as a new free parameter in the training process. A unified framework for training KOL algorithm is proposed, where the kernel weight and kernel width are updated adaptively at the same time. Furthermore, to make the system convergent in the sense of external disturbance, a dead zone scheme is applied to determine the learning rate. For the sparsification procedure, a new rule based on the cumulative coherence of a dictionary is proposed. The dictionary is selected online by minimizing its cumulative coherence. By applying the proposed training method and sparsification rule, we propose a kernel algorithm with adaptive kernel width (KAW) for online learning.

The organization of this paper is as follows. In Section 2, we introduce some fundamentals of kernel methods and KOL algorithms. In Section 3, the unified framework for updating the kernel weight and kernel width is proposed. The adaptive training method is detailed and the theoretical analysis of its convergence is presented. Section 4 presents the sparsification rule based on the online minimization of cumulative coherence. In Section 5, simulation results of several examples are presented and finally conclusion is given in Section 6.

2. Kernel online learning algorithms

The fundamental idea of kernel methods is the use of Mercer kernel function. It transforms the feature vectors from the low input feature space to the higher dimensional RKHS. The nonlinear mapping in the feature space results in a linear solution for kernel methods.

For a training sequence $\{\boldsymbol{u}(j), d(j)\}_{j=1}^{t}$, the estimated function $f_t(\cdot)$ in kernel methods to model the output can be expressed as

$$f_t(\cdot) = \sum_{j=1}^{t} \alpha_j(t)\kappa(\boldsymbol{u}(j), \cdot) \tag{1}$$

where $\kappa(\boldsymbol{u}(j), \cdot)$ is a Mercer kernel, $\boldsymbol{u}(j)$ and $\alpha_j(t)$ is the kernel center and coefficient respectively. The estimated output function is a linear combination of series kernel functions. This leads to a simplicity in the structure and computational complexity, which makes it suitable for real-time online learning. The linearity in its solution form also allows many linear methods able to be extended into the RKHS for its training and updating.

2.1. Online sparsification rules

Without any regularization for the structure of kernel methods as in (1), the kernel function number increases by one when a new training sample arrives. The continuous growing number of kernel functions make the computational complexity expensive which grows cubically with the number of training samples. In online learning context, the computational cost and structure complexity of KOL algorithms should be affordable to make it work in real time. An effective way to solve this problem is to select a representative dictionary, which is also known as "sparsification". In the kernel algorithm featured by a dictionary $\{(c_j(t)\}_{j=1}^{m}$ with m members at $t(th)$ training iteration, the estimation function $f_t(\cdot)$ becomes

$$f_t(\cdot) = \sum_{i=1}^{m} \alpha_i(t)\kappa(\boldsymbol{c}_i(t), \cdot). \tag{2}$$

The kernel function number is reduced to m, $(m \ll t)$ in the sparse representation (2). Straightforwardly, as in traditional kernel methods like support vector machines or support vector regression, the sparsification operation would not only reduce the computational cost heavily but also increase the generalization performance [21–24].

The sparsification is a process to select the novel and informative training samples as the dictionary members. In the last decade, there have been many criteria proposed for this purpose, including the approximate linear dependence (ALD) criterion [2], the coherence-based criterion [3], the novelty (NC) criterion [12] and the surprise criterion [11], etc.

2.2. Motivation of the paper

In the training process of KOL algorithms, we consider the two most important issues: the structure complexity regularization and free parameters updating. The former issue aims to reduce the computational cost as well as increase the generalization performance and the later one is to make the kernel algorithm able to reach the optimal steady state with a fast convergence speed. The KLMS or KNLMS algorithms are type of gradient based algorithms. In the learning rules, a fixed learning rate η is adopted for the whole training iterations. However, a fixed learning rate is not the optimal choice for leaning algorithms. Usually a small learning rate need to be chosen to make the algorithm convergent and a subsequent problem is that the small learning rate would also make the learning process very slow, especially at the beginning part [25]. When the system reaches its optimal status the system may suffer from over-fitting problem because the training error mainly comes from the external disturbance.

In the recent KOL algorithms, the kernel width σ is a fixed value and determined manually by prior knowledge or exhausting search in preliminary experiments. The kernel width is not automatically adapted by the training data. Consider the commonly used Mercer kernel: the Gaussian kernel

$$\kappa(\boldsymbol{u}(i), \boldsymbol{u}(j)) = \exp\left(-\frac{\|\boldsymbol{u}(i) - \boldsymbol{u}(j)\|^2}{2\sigma^2}\right) \tag{3}$$

kernel width plays an important role in the tradeoff between presenting the locality of data and smoothness of learning process [16]. Thus to overcome the drawback of using a fixed kernel width, we propose a new LMS algorithm based kernel online learning algorithm which combines the kernel weight updating and kernel width updating together. The training method adopts an adaptive learning rate by a dead zone scheme to assure its convergence. The existing sparsification rules all require one or two parameters to determine the sparsity of the dictionary which are dependent on the prior knowledge of the training data. Instead, a new sparsification rule is proposed based on the cumulative coherence of the dictionary. A compact dictionary with predefined size can be selected adaptively without any prior information of the training data.