Expert Systems with Applications 41 (2014) 4349–4359

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An information theoretic sparse kernel algorithm for online learning

Haijin Fan [a,*], Qing Song [a], Zhao Xu [b]

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

ARTICLE INFO

Keywords:
Kernel methods
Information theoretic
Sparsification
Online learning
Mutual information

ABSTRACT

Kernel-based algorithms have been proven successful in many nonlinear modeling applications. However, the computational complexity of classical kernel-based methods grows superlinearly with the increasing number of training data, which is too expensive for online applications. In order to solve this problem, the paper presents an information theoretic method to train a sparse version of kernel learning algorithm. A concept named instantaneous mutual information is investigated to measure the system reliability of the estimated output. This measure is used as a criterion to determine the novelty of the training sample and informative ones are selected to form a compact dictionary to represent the whole data. Furthermore, we propose a robust learning scheme for the training of the kernel learning algorithm with an adaptive learning rate. This ensures the convergence of the learning algorithm and makes it converge to the steady state faster. We illustrate the performance of our proposed algorithm and compare it with some recent kernel algorithms by several experiments.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Kernel methods have been proven to be successful tools in modeling of nonlinear systems. They are based on the use of Mercer kernels which map the input feature vector in a low dimensional space into a high or even infinite dimensional space, so that a wide range of nonlinear problems can find their optimal or suboptimal solutions in the high dimensional feature space (Schölkopf, Herbrich, & Smola, 2001). Due to their solid mathematical foundation, kernel methods become very popular in practical applications. The widely used approaches include support vector machines (SVMs) (Chen, Li, Wei, Xu, & Shi, 2011; Fu & Lee, 2012; Ha, Wang, & Chen, 2012), Gaussian Process (GP) theory (Deisenroth, Rasmussen, & Peters, 2009; Huang, 2011; Srijith, Shevade, & Sundararajan, 2013), kernel least mean square (KLMS) algorithm and its extensions (Príncipe, Liu, & Haykin, 2011). According to the kernel representer theorem (Schölkopf et al., 2001), the number of kernel functions will grow linearly with the number of training samples, which leads to a superlinearly growing computational cost. Most of the current kernel methods are for off-line learning, the training samples are available in advance and the computational complexity is very high. However, online version of kernel learning algorithms require a higher computational speed.

Recently, several online kernel algorithms have been proposed (Engel, Mannor, & Meir, 2004; Fan & Song, 2014; Li et al., 2013; Richard, Bermudez, & Honeine, 2009; Van Vaerenbergh, Lazaro-Gredilla, & Santamaria, 2012). These algorithms applied the mean square error as their loss function and searched their solutions in the reproducing kernel Hilbert space (RKHS). To reduce the number of kernel functions and the computational complexity of the algorithms, different sparsification criteria have been investigated to select an informative and compact dictionary. A truncation method was applied to delete the earliest training observations since the approximation error without them vanished exponentially (Kivinen, Smola, & Williamson, 2004). The approximate linear dependence (ALD) (Engel et al., 2004) and coherence-based criterion (Richard et al., 2009) were discussed and incorporated into the kernel learning algorithm. These criteria at some extent successfully selected the most representable samples as the dictionary members in a feature dependence perspective. On the other hand, from an information theoretic point of view, relative information measures have been very popular in independent and novel data detection. Mutual information based methods were used for clustering (Song, 2005), variable dependence detection and feature selection (Roberto, 2010; Oveisi, Oveisi, Erfanian, & Patras, 2012). Kullback–Leibler divergence measure was applied for novelty detection (Filippone & Sanguinetti, 2010). These information theoretic methods work efficiently according to the distribution of input features or variables. However, they are based on the characteristics of the system input, without considering the

* Corresponding author. Tel.: +65 85525016.
E-mail address: hfan1@e.ntu.edu.sg (H. Fan).

http://dx.doi.org/10.1016/j.eswa.2014.01.010
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

reliability of the estimated system output. In a different perspective, an information theoretic approach based on the surprise measure was proposed for different kinds of sparse kernel adaptive filters (Liu, Park, & Príncipe, 2009). The surprise measure was defined as the negative log likelihood of the joint distribution of the system input and output. Those training samples with relatively high surprise are considered as dictionary members.

Besides the computational complexity, the weight convergence and generalization ability are two critical challenges for online learning algorithms. These issues in neural learning system were examined in Song (2010, 2011). Approaches that guarantee the weight convergence in the learning system make the system able to find its optimal or suboptimal solution fast. The problem of overfitting which reduces the generalization ability of the algorithms can be generally avoided by using the stopping training scheme or adding a regularization term in the loss function. As in SVMs, a reasonable weight norm constraint was included to make it more generalized for nonseparable data (Abe, 2010). Stopping training the redundant data is also a way to overcome overfitting problem. Heuristically, a small training error means the existing system can fit the training sample well and the corresponding sample carries little useful information for the system updating. The learning system should stop learning it. According to this, a dead zone scheme was proposed in neural networks for robust learning with a time-varying learning rate (Song, 2011). It was shown that an adaptive learning rate could guarantee the convergence of the weight in the sense of the Lyapunov function of the weight error vector (Song, 2010, 2011).

Kernel methods could be considered as one layer neural networks. The approaches and methods developed in neural networks can be applied to examine kernel methods (Mackenzie & Tieu, 2004). In this paper, an online robust kernel learning algorithm is proposed. To reduce the computational cost, an information theoretic sparsification method is developed to reduce the number of kernel functions. Furthermore, an adaptive and robust learning scheme is applied to improve the generalization ability of the algorithm. We associate the sparsification method with the adaptive learning scheme and derive an information theoretic sparse kernel algorithm (ITSKL) for online applications. The main contributions of the paper are two folds. Firstly, the problem of weight convergence in online kernel learning algorithms is critically important. On most of the existing kernel methods, this issue is rarely investigated. To ensure the weight convergence, we develop an adaptive training methods for the kernel weight updating. Secondly, a new sparsification method based on the instantaneous mutual information is investigated to reduce the number of kernel functions. The new information theoretic method is with less computational complexity and has a good performance.

The paper is organized as follows. In Section 2, some fundamental ideas of kernel methods are introduced. In Section 3, the information theoretic sparsification method is discussed. Section 4 gives the adaptive and robust learning scheme based on the dead zone scheme. Section 5 shows the experiment results and conclusion is given in Section 6.

2. Nonlinear regression with kernel

In machine learning, online learning is an important learning scenario. Usually in the learning settings for batch learning or off-line learning, the whole training samples are available at hand in advance and the learning systems are trained by a batch of training samples at every iteration. However, in online learning context, a sequence of training samples are presented to the learning system one by one at each iteration. Considering the supervised learning task with the recorded training input–output pairs for a unknown system:

$$S(t) = \{(\boldsymbol{u}_1, d_1), (\boldsymbol{u}_2, d_2)), \ldots, (\boldsymbol{u}_n, d_n)\} \tag{1}$$

where $\boldsymbol{u}_t \in \mathbb{R}^d$ is the input feature vector and $d_t \in \mathbb{R}$ is the desired output at the t(th) iteration. The goal of the online learning algorithm is to find a linear or nonlinear function to predict the estimated output $y_{t+1} = f_t(\boldsymbol{u}_{t+1})$ that best fits the ideal output d_{t+1} given the input feature vector $\boldsymbol{u}_{t+1}$. Online learning algorithms obtain the current prediction model $f_t(\cdot)$ without retraining it from scratch. Instead, they update the corresponding parameters and train it from the previous estimation $f_{t-1}(\cdot)$. Compared with batch learning or off-line learning algorithms, online learning algorithms are applicable to real-time applications and thus require much more computationally efficient training methods as well as less expensive storage burden to store the training sample information.

2.1. Problem formulation

In kernel based regression, a positive definite kernel is used to compute the inner product of features in the higher dimensional space. Let $\mathcal{U} \times \mathcal{U} \to R$ denote the kernel mapping and $\mathcal{H}$ be the corresponding RKHS. Given an input feature vector $\boldsymbol{u}_i$ and its desired output d_i, the problem is to find a function $f(\cdot)$ to reconstruct the corresponding output $f(\boldsymbol{u}_i) = \langle \psi(\cdot), k(\cdot, \boldsymbol{u}_i) \rangle$ such that the loss function becomes minimized

$$min_{f \in \mathcal{H}} R_{emp}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \tag{2}$$

where $R_{emp}(f)$ represents the empirical loss function and $\|f\|_{\mathcal{H}}^2$ is the regularization term measuring the complexity of the system with $\lambda > 0$ (Kivinen et al., 2004). By virtue of the representer theorem, the function $f(\cdot)$ in (2) can be expressed as a linear combination of kernel functions

$$f(\cdot) = \sum_{j=1}^{m} \alpha_j \kappa(\cdot, \boldsymbol{u}_j) \tag{3}$$

where $\kappa(\cdot, \boldsymbol{u}_j)$ is a kernel function centered at the input feature vector $\boldsymbol{u}_j$ and α_j is the weight coefficient. As a result, the function can be estimated implicitly by the input feature vectors in $\mathcal{U}$ with a Mercer kernel function. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top \in \mathbb{R}^{m \times 1}$, $\boldsymbol{d} = (d_1, \ldots, d_n)^\top \in \mathbb{R}^{n \times 1}$ and $\boldsymbol{K} \in \mathbb{R}^{n \times m}$, respectively stand for the weight vector, the desired output, and the Gram matrix where $\boldsymbol{K}_{ij} = \kappa(\boldsymbol{u}_i, \boldsymbol{u}_j)$. The problem can be formulated as a least square error problem $min_{\boldsymbol{\alpha}} \|\mathbf{d} - \mathbf{K}\boldsymbol{\alpha}\|^2$. The solutions of this problem are twofolds: the choice of $\boldsymbol{u}_j$ and its weight coefficient α_j. The kernel function is centered at the input feature vector $\boldsymbol{u}_j$, which specifies the characteristics of the estimated function. On the one hand, the selection of a compact dictionary $D_t = \left\{(\boldsymbol{u}_j, y_j)_{j=1}^m\right\}$ is needed. On the other hand, $\boldsymbol{\alpha}$ can be found by solving the linear equation $\boldsymbol{d} = \boldsymbol{K}\boldsymbol{\alpha}$. Our proposed methods for these two problems are presented in Section 3 and Section 4 respectively.

3. Online sparsification method

The bottleneck of the kernel algorithm for online learning is its computational complexity increase with the growing number of training samples. As in the kernel model (3), its computational cost is very expensive, usually with a scale of $\mathcal{O}(n^3)$ in time and a memory scale of $\mathcal{O}(n^2)$, where n is the number of training samples. In real time applications when training samples are very large, the kernel method without sparse representation will encounter a big problem: it will be very slow and need a large memory space to store the sample information. Many methods were proposed to learn a sparse dictionary to reduce the number of kernel functions. An improved Lasso method with efficient computation approaches

Download English Version:

https://daneshyari.com/en/article/386594

Download Persian Version:

https://daneshyari.com/article/386594

Daneshyari.com