Neurocomputing 149 (2015) 966–974

Contents lists available at ScienceDirect

Neurocomputing

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

The graph based semi-supervised algorithm with ℓ^1-regularizer

Ling Zuo [a,b], Luoqing Li [a,*], Chen Chen [a]

[a] Faculty of Mathematics and Statistics, Hubei University, Wuhan 430062, China
[b] School of Science, Hubei University of Technology, Wuhan 430068, China

ARTICLE INFO

Article history:
Received 29 January 2014
Received in revised form
8 April 2014
Accepted 21 July 2014
Communicated by Feiping Nie
Available online 2 August 2014

Keywords:
Graph-based semi-supervised learning
ℓ^1−regularizer
Excess misclassification error
Hypothesis error
Manifold error

ABSTRACT

In this paper a new graph-based semi-supervised algorithm for regression problem is proposed. An excess generalization error bound is established. It evaluates the learning performance of the proposed method and has a fast convergence rate with $O(l^{\epsilon-1})$ decay. An example is given to show that the proposed method uses a small portion of the labeled and unlabeled data to represent the target function, which illustrates the sparsity of the algorithm, and can efficiently reduce the computational complexity of the semi-supervised learning. Moreover, some experiments are performed to validate the sparsity and learning performance of the formulation.

1. Introduction

Semi-supervised learning (SSL), as a powerful tool to learn from a small number of labeled data and a large number of unlabeled data, has been of growing interest in the machine learning research [1–4]. There are two typical SSL approaches: learning with the cluster assumption [5] and learning with the manifold assumption [6–9]. When assuming that the data is embedded into a low-dimensional manifold, the graph-based method seems more effective as the unlabeled data can be used to uncover the intrinsic manifold structures. Solid theoretical foundations have been laid for a large family of graph-based SSL algorithms, such as Laplacian support vector machine (LapSVM) and Laplacian regularized least squares (LapRLS). However, most of these algorithms have solutions that involve kernel evaluation on all the labeled and unlabeled examples, which may result in a much slower computation dealing with the huge set of the unlabeled data in real applications.

To address the above issue, previous discussions have done to realize sparse graph-based SSL. For example, Tsang et al. [7] proposed a sparse graph-based SSL algorithm by introducing the sparsified manifold regularizer. Sun et al. [10] caught the sparsity of the algorithm based on the manifold-preserving graph which reduces the number of vertices in the graph. Other examples include [11] and [12]. These algorithms can be unified in a Tikhonov regularization scheme in a reproducing kernel Hilbert space (RKHS) with a Mercer kernel K. Such a kernel has to be continuous, symmetric and positive semi-definite.

In this paper we proposed a novel graph-based semi-supervised algorithm for regression problems. Unlike the Tikhonov regularization scheme in a RKHS, the algorithm considered here takes the form of regularization scheme with both the ℓ^1-regularizer and manifold regularizer that stated in a data dependent hypothesis space. Note that the kernel here is not necessarily symmetric or semi-definite. This leads to a wider selection of the kernel which offers more flexibility. Explicit examples of this general kernel can be found in [13,14].

The main results of the paper are as follows. One is that we derive the generalization error bound for the proposed algorithm and show that it has $O(l^{\epsilon-1})$ decay, where $\epsilon>0$ is a small parameter tends to zero. This convergence rate is faster than the sparse semi-supervised method in [11] with the order of $O(l^{-1/2})$. It is also faster than the supervised coefficient regularization methods, e.g., $O(l^{-1/5})$ in [15], $O(l^{-1/3})$ in [13], and $O(l^{-1/2})$ in [14]. The main difficulty in the error analysis is the data dependent nature of the kernel-based hypothesis space and the extra manifold regularization term. Hence, a new error decomposition technique is exploited by means of introducing the additional hypothesis error and manifold error. The other contribution is that we study the sparsity of the algorithm based on our error analysis. An example is given to show that the proposed algorithm is able to use a small portion of the labeled and unlabeled samples to represent the target function. This is computationally desirable, because for semi-supervised learning usually a considerably large number of unlabeled examples are available.

* Corresponding author.
E-mail addresses: zuolingcc@gmail.com (L. Zuo), lilq@hubu.edu.cn (L. Li).

http://dx.doi.org/10.1016/j.neucom.2014.07.037
0925-2312/

The remainder of the paper is organized as follows. In Section 2, we present the new graph-based semi-supervised method. Section 3 includes the main results on error analysis and the sparsity of the algorithm. In Section 4 the generalization error is estimated based on the empirical covering numbers. In Section 5 the sparsity of the algorithm is investigated based on our error analysis. An empirical study is given in Section 6. We conclude the paper in Section 7.

2. The new graph-based semi-supervised algorithm with the ℓ^1-regularizer

In this section, we first provide the necessary background of the semi-supervised regression problem and then present the new graph-based semi-supervised algorithm.

2.1. Preliminaries

In a regression problem, we work with an input space X and an output space Y. Let $X\subset\mathbb{R}^d$ be a compact metric space and $Y=[-M,M]$. A function $f: X\to Y$ makes a prediction of $y\in Y$ at $x\in X$ by $f(x)$. The prediction accuracy may be measured by the least-square loss $(y-f(x))^2$. Let ρ be an underlying probability distribution on $Z=X\times Y$. The prediction ability of f is quantitatively measured by the generalization error

$$\mathcal{E}(f)=\int_Z (y-f(x))^2\, d\rho.$$

Decompose ρ into the marginal distribution ρ_X on X and the conditional distribution $\rho(y|x)$ at $x\in X$. The function minimizing $\mathcal{E}(f)$ is called the regression function given by

$$f_\rho(x)=\int_Y y\, d\rho(y|x),\quad x\in X.$$

Since ρ is usually unknown, f_ρ cannot be obtained directly. We can learn f_ρ from samples. Kernel method is an important tool in learning theory. A well studied kernel-based algorithm for the regression problem is the least-square regularization scheme. If $K: X\times X\to\mathbb{R}$ is a continues positive semi-definite kernel and $(\mathcal{H}_K,\|\cdot\|_K)$ is the associated RKHS which is data independent, then the scheme is given by

$$f_{\mathbf{z},\lambda}=\arg\min_{f\in\mathcal{H}_K}\{\mathcal{E}_{\mathbf{z}}(f)+\lambda\|f\|_K^2\},\tag{1}$$

where $\mathcal{E}_{\mathbf{z}}(f)=1/l\sum_{i=1}^{l}(y_i-f(x_i))^2$ is the empirical risk and $\lambda>0$ is a regularization parameter. Mathematical analysis of learning algorithm (1) has been well understood [16].

In the semi-supervised model, a learner obtains a labeled data set $\mathbf{z}=\{(x_i,y_i)\}_{i=1}^{l}$ and an unlabeled data set $\mathbf{x}_1=\{x_j\}_{j=l+1}^{j=l+u}$. It aims at using a large amount of unlabeled data, together with a small set of labeled data, to built better predictors f. Throughout the paper we assume that the labeled data are drawn independently according to the measure ρ while the unlabeled data are generated independently by the marginal distribution ρ_X.

When the data lie on a low-dimensional manifold, the graph-based method seems more effective for semi-supervised learning and many approaches have been proposed. One of the most successful works is the manifold regularization framework proposed in [6] due to the convexity of its optimization problem, out-of-sample prediction and solid theoretical foundations. They introduced the LapRLS which was formulated as

$$\min_{f\in\mathcal{H}_K}\frac{1}{l}\sum_{i=1}^{l}(y_i-f(x_i))^2+\lambda_1\|f\|_K^2+\frac{\lambda_2}{(l+u)^2}\hat{f}^TL\hat{f}$$

where $\hat{f}=(f(x_1),\ldots,f(x_{l+u}))^T$. λ_1 and λ_2 are nonnegative regularization parameters. $L=D-W$ is the unnormalized graph Laplacian, where D is a diagonal matrix with diagonal entries $D_{ii}=\sum_{j=1}^{l+u}W_{ij}$. The weight W_{ij} is given by a similar function $W(x_i,x_j)$. The more similar x_i and x_j, the larger W_{ij} should be. Note that $\hat{f}^TL\hat{f}=\frac{1}{2}\sum_{i,j=1}^{l+u}(f(x_i)-f(x_j))^2W_{ij}$.

In this framework, the underlying geometric structure of the marginal distribution is estimated from the unlabeled data and is incorporated into a RKHS to form a regularization term (a.k.a. manifold regularizer). This regularizer ensures that the learned decision function is smooth along the manifold. Recent work [17] also shows that the manifold regularizer is useful for the feature selection.

2.2. The new graph-based semi-supervised algorithm with the ℓ^1-regularizer

In this paper we propose a new graph-based semi-supervised formulation which abandons the symmetry (and of course positive semi-definiteness) of the kernel, and considers a regularization scheme with ℓ^1-regularizer stated in a data dependent hypothesis spaces. Here a kernel function $K: X\times X\to\mathbb{R}$ is a continuous function. The hypothesis space depends on the sample $\mathbf{z}$ and is defined as

$$\mathcal{H}_{K,\mathbf{z}}=\left\{\sum_{i=1}^{l+u}\alpha_iK_{x_i}:\alpha_i\in\mathbb{R}\right\},$$

where $K_t(\cdot)=K(\cdot,t)$. The proposed algorithm is given by

$$f_{\mathbf{z},\lambda}=\arg\min_{f\in\mathcal{H}_{K,\mathbf{z}}}\left\{\mathcal{E}_{\mathbf{z}}(f)+\lambda_1\Omega_{\mathbf{z}}(f)+\frac{\lambda_2}{(l+u)^2}\hat{f}^TL\hat{f}\right\},\tag{2}$$

where

$$\Omega_{\mathbf{z}}(f)=\sum_{i=1}^{l+u}|\alpha_i|\quad\text{for } f=\sum_{i=1}^{l+u}\alpha_iK_{x_i}.$$

Assume that $0\le W_{ij}\le\omega$ with $\omega\ge 0$.

This nonsymmetric kernel appears naturally in the study of dual wavelets or frames in wavelet analysis [18,19]. It has the flexibility of having good representation for $f_{\mathbf{z},\lambda}$ while keeping strong approximation ability. The ℓ^1-regularizer often leads to some sparse properties. Note that when $u=0$ and $\lambda_2=0$, algorithm (2) reduces to the proposed method in [13]. In this paper, we are interested in how fast $f_{\mathbf{z},\lambda}$ approximates f_ρ as the number of examples increases. Moreover, based on our error analysis the sparsity of algorithm (2) will be studied.

3. Main results

In this section, learning rates will be provided in terms of properties of the input space X, the measure ρ, and the kernel K. Besides, an example is given to show the sparsity of the proposed algorithm.

Denote by $\mathcal{N}(X,r)$ the covering number of X. Throughout the paper we shall assume

$$\mathcal{N}(X,r)\le C_\eta\left(\frac{1}{r}\right)^\eta,\quad\forall 0<r\le 1\tag{3}$$

for some $\eta>0$ and $C_\eta>0$.

Definition 3.1. A probability measure ρ_X on X is said to satisfy condition L_τ with $\tau>0$ if there exists some $C_\tau>0$ such that

$$\rho_X(\{u\in X:|u-x|\le r\})\ge C_\tau r^\tau,\quad\forall x\in X,\ 0<r\le 1.$$