Neurocomputing 149 (2015) 180–186

Contents lists available at ScienceDirect

Neurocomputing

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

# Semi-supervised extreme learning machine with manifold and pairwise constraints regularization

Yong Zhou, Beizuo Liu, Shixiong Xia*, Bing Liu

*School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China*

ARTICLE INFO

*Article history:*
Received 8 August 2013
Received in revised form
18 December 2013
Accepted 19 January 2014
Available online 7 October 2014

*Keywords:*
Semi-supervised learning
Extreme learning machine (ELM)
Pairwise constraints
Regularization

ABSTRACT

Traditional kernel-based semi-supervised learning (SSL) algorithms usually have high computational complexity. Moreover, few SSL methods have been proposed to utilize both the manifold of unlabeled data and pairwise constraints effectively. In this paper, we first construct a unified SSL framework to combine the manifold regularization and the terms based on the pairwise constraints for semi-supervised classification tasks. Motivated by the effectiveness of extreme learning machine (ELM), we further utilize ELM to approximate the established kernel-based SSL framework. Finally, we present a fast semi-supervised extreme learning machine with manifold regularization and pairwise constraints. Experimental results on a variety of real-world data sets demonstrate the effectiveness of the proposed fast SSL algorithm.

© 2014 Elsevier B.V. All rights reserved.

## 1. Introduction

In many practical applications, one often faces a lack of sufficient labeled data. One way to address the problem of the small-size samples is to utilize a large amount of unlabeled data by semi-supervised learning methods. A large number of SSL approaches have been proposed over the years, such as self-training, co-training, transductive support vector machines (TSVMs) and graph-based methods. Among them, kernel functions are often used to enhance the performance of SSL. However, SSL methods based on kernel functions generally have high computational complexity. How to make these SSL methods applicable to large-scale data sets becomes a very challenging task. Recently, Extreme learning machine (ELM) has recently become an interesting topic because of its fast learning capacity [1–9]. Thus, it is natural to introduce it into kernel-based SSL methods, which not only provides an approximate method for traditional SSL methods, but also extends ELM to the semi-supervised scenario.

Now, many regularization frameworks have been designed by using manifold regularization terms based on the manifold assumption, that is, the samples in the local region should have similar labels. Belkin et al. [10] first proposed a general manifold regularization (MR) framework developed in the setting of Reproducing Kernel Hilbert Spaces (RKHS). To measure the smoothness of functions on data manifolds, the MR framework added an additional penalty term to the traditional regularization. By exploiting the intrinsic structure of data, such a term can enhance the smoothness of decision functions and further improve the performance of learning algorithms. Based on the MR framework, the Discriminatively Regularization Least Square Classification (DRLSC) method built the penalty term on manifolds by integrating both discriminative and geometrical information in each local region [11]. These frameworks can handle semi-supervised learning problems well, but they do not utilize pairwise constraints effectively. Moreover, the high computational complexities of algorithms from these frameworks limit the application of these methods.

Recently, some semi-supervised ELM methods have been proposed to improve the original ELM model. Liu et al. [12] developed a semi-supervised ELM (SELM) model by introducing the manifold regularization term. Li et al. [13] proposed a new regularization classification method (NRCM) by constructing the intra-class and inter-class regularization terms. Although these methods can utilize unlabeled data, they did not exploit the information from pairwise constraints of unlabeled data, which could be available in practical applications. Moreover, the complexity of these models was not controlled, that is, the penalty norm of the function in the ambient space was not incorporated into these models. Undoubtedly, this could have a bad influence on the performance of algorithms.

In this paper, we are particularly interested in how to incorporate pairwise constraints to the traditional manifold regularization framework and how to enhance the effectiveness of traditional kernel-based learning algorithms. We first construct a unified SSL

* Corresponding author.
*E-mail address:* xiasx@cumt.edu.cn (S. Xia).

http://dx.doi.org/10.1016/j.neucom.2014.01.073
0925-2312/© 2014 Elsevier B.V. All rights reserved.

framework to combine the manifold assumption and the pairwise constraints. Then we approximately solve the proposed model by virtue of output functions in the ELM feature space and present a fast semi-supervised learning algorithm based on ELM (SSL-ELM).

The remainder of this paper is organized as follows: Section 2 introduces the regularization technology. Section 3 presents an SSL framework in detail and proposes a fast SSL algorithm based on ELM. Experimental results on a synthetic data set and several real-world data sets are reported in Sections 4 and Section 5 is conclusions.

## 2. Extreme learning machine

The output function of ELM for generalized SLFNs in the case of one output node is

$$f_L(\boldsymbol{x}) = \sum_{i=1}^{L} \beta_i h_i(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\beta} \tag{1}$$

where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_L]^T$ is the vector of the output weights between the hidden layer of $L$ nodes and the output node, and $\boldsymbol{h}(\boldsymbol{x}) = [h_1(\boldsymbol{x}), \ldots, h_L(\boldsymbol{x})]$ is the output (row) vector of the hidden layer with respect to the input $\boldsymbol{x}$. In fact, $\boldsymbol{h}(\boldsymbol{x})$ maps the data from the $d$-dimensional input space to the $L$-dimensional hidden-layer feature space (ELM feature space) $\boldsymbol{H}$. ELM is to minimize the training error as well as the norm of the output weights [14,15]

$$min_{\beta}\ \frac{C}{2}\|H\boldsymbol{\beta} - \boldsymbol{T}\|_F^2 + \frac{1}{2}\|\boldsymbol{\beta}\|^2 \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\boldsymbol{H}$ is the hidden-layer output matrix denoted by

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{h}(\boldsymbol{x}_1) \\ \boldsymbol{h}(\boldsymbol{x}_2) \\ \vdots \\ \boldsymbol{h}(\boldsymbol{x}_n) \end{bmatrix} = \begin{bmatrix} h_1(\boldsymbol{x}_1) & \ldots & h_L(\boldsymbol{x}_1) \\ h_1(\boldsymbol{x}_2) & \ldots & h_L(\boldsymbol{x}_2) \\ \vdots & \vdots & \vdots \\ h_1(\boldsymbol{x}_n) & \ldots & h_L(\boldsymbol{x}_n) \end{bmatrix}. \tag{3}$$

Similar to SVM, to minimize the norm of the output weights $\|\beta\|$ is actually to maximize the distance of the separating margins of the two different classes in the ELM feature space: $2/\|\beta\|$, which actually controls the complexity of the function in the ambient space.

## 3. The SSL framework with the manifold assumption and pairwise constraints

### 3.1. Manifold and pairwise constraints regularization

In graph-based SSL methods, the manifold assumption is widely used. There is a probability distribution $\boldsymbol{P}$ on $\boldsymbol{X} \times \mathbb{R}$ according to which examples are generated for function learning. Labeled examples are $(\boldsymbol{x}, \boldsymbol{y})$ pairs generated from $\boldsymbol{P}$. Unlabeled examples are simply $\boldsymbol{x} \in \boldsymbol{X}$ drawn according to the marginal distribution $\boldsymbol{P}_{\boldsymbol{x}}$ of $\boldsymbol{P}$. Previous studies have shown that there may be a connection between the conditional and margin distributions. Thus, knowledge of the margin $\boldsymbol{P}_{\boldsymbol{x}}$ can be exploited for better function learning. Specifically, if two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in X$ are close in the intrinsic geometry of $\boldsymbol{P}_{\boldsymbol{x}}$, then the conditional probabilities $\boldsymbol{P}(\boldsymbol{y}|\boldsymbol{x}_1)$ and $\boldsymbol{P}(\boldsymbol{y}|\boldsymbol{x}_2)$ are similar, where $\boldsymbol{y} \in \{1, \ldots, m\}$ is the class label. Thus, the conditional probability distribution varies smoothly along the geodesics in the intrinsic geometry of $\boldsymbol{P}_{\boldsymbol{x}}$. This is usually referred to as manifold assumption [16].

For a Mercer kernel $K$: $\boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}$, there is an associated RKHS $\mathscr{H}_K$ of functions $\boldsymbol{X} \to \mathbb{R}$ with the corresponding norm $\|\ \|_{\mathscr{H}}$. Given a set of $l$ labeled examples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{l}$ and a set of $u$ unlabeled examples $\{\boldsymbol{x}_j\}_{j=l+1}^{l+u}$, In the manifold regularization framework, an unknown function is estimated by minimizing [10]

$$\boldsymbol{f}^* = \underset{\boldsymbol{f} \in \mathscr{H}}{\operatorname{argmin}} \left[ \frac{1}{l} \sum_{i=1}^{l} V(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{f}) + \gamma_A \|\boldsymbol{f}\|_{\mathscr{H}}^2 + \gamma_I \|\boldsymbol{f}\|_{\boldsymbol{I}}^2 \right] \tag{4}$$

where $V$ is some loss function, i.e., the squared loss $(\boldsymbol{y_i} - f(\boldsymbol{x_i}))^2$ for RLS or the hinge loss function $\max[0, 1 - \boldsymbol{y_i} f(\boldsymbol{x_i})]$ for SVM, $\|\boldsymbol{f}\|_{\mathscr{H}}^2$ is the RKHS norm penalty and represents the complexity of the function in RKHS $H_K$ and $\|\boldsymbol{f}\|_I^2$ is a smoothness penalty corresponding to the sample probability distribution. $\gamma_A$ controls the complexity of the function in the ambient space and $\gamma_1$ controls the complexity of the function in the intrinsic geometry of sample probability distribution.

When we consider the case that the support of $\boldsymbol{P}_{\boldsymbol{x}}$ is a compact submanifold $\mathscr{M} \subset \mathbb{R}^{\mathrm{d}}$, a natural choice for $\|\boldsymbol{f}\|_I$ is $\int_{x \in \mathscr{M}} \|\nabla_{\mathscr{M}} \boldsymbol{f}\|^2 d\boldsymbol{P}_{\boldsymbol{x}}(\boldsymbol{x})$ [14], where $\nabla_{\mathscr{M}}$ is the gradient of $\boldsymbol{f}$ along the manifold and the integral is taken over the distribution $\boldsymbol{P}_{\boldsymbol{x}}$. In most applications the marginal $\boldsymbol{P}_{\boldsymbol{x}}$ is unknown. Therefore we must attempt to get empirical estimates of $\boldsymbol{P}_{\boldsymbol{x}}$ and $\|\cdot\|_I$. In order to model the geometrical structure of $\mathscr{M}$, we construct a nearest-neighbor graph $\boldsymbol{G}$ and define a weighted matrix $W$ on the graph. Define $\boldsymbol{L_G} = \boldsymbol{D} - \boldsymbol{W}$, where $D$ is a diagonal matrix whose entries are column (or row) sums of $W$, that is $\boldsymbol{D}_{ii} = \sum_{j=1}^{l+u} \boldsymbol{w}_{ij}$. $L_G$ is called graph Laplacian [14]. By spectral graph theory, $\|\boldsymbol{f}\|_I^2$ can be discretely approximated as follows:

$$\begin{aligned} \|\boldsymbol{f}_I^2\| &= \frac{1}{2(u+l)^2} \sum_{i,j=1}^{l+u} \|f(\boldsymbol{x_i}) - f(\boldsymbol{x_j})\|^2 \boldsymbol{w}_{ij} \\ &= \frac{1}{(u+l)^2} \left( \sum_{i=1}^{l+u} f(\boldsymbol{x_i})^2 \boldsymbol{D}_{ii} - \sum_{i=1,j=1}^{l+u} f(\boldsymbol{x_i}) f(\boldsymbol{x_j}) \boldsymbol{w}_{ij} \right) \\ &= \frac{1}{(u+l)^2} \left( \boldsymbol{f}^T \boldsymbol{D} \boldsymbol{f} - \boldsymbol{f}^T \boldsymbol{W} \boldsymbol{f} \right) = \frac{1}{(u+l)^2} \boldsymbol{f}^T \boldsymbol{L_G} \boldsymbol{f}, \end{aligned} \tag{5}$$

where the normalizing coefficient $1/(l+u)^2$ is the natural scale factor for the empirical estimate of the Laplace operator.

In order to utilize pairwise constraints, we first construct two weighted matrices based on pairwise constraints as follows:

$$\boldsymbol{W_{m,ij}} = \begin{cases} 1 & if\ x_i, x_j \in ML \\ 0 & \text{Otherwise} \end{cases} \tag{6}$$

and

$$\boldsymbol{W_{c,ij}} = \begin{cases} 1 & \text{if } x\, x_i, x_j \in CL \\ 0 & \text{Otherwise} \end{cases} \tag{7}$$

where $ML$ represents the set of the must-link pairwise constraints and $CL$ represents the set of the cannot-link pairwise constraints. Thus, we actually construct the intra-class graph $\boldsymbol{G_m}$ and the inter-class graph $\boldsymbol{G_c}$. Then, we define a measure to characterize the intra-class compactness from the intra-class graph

$$\begin{aligned} \boldsymbol{S_m} &= \frac{1}{2} \sum_{i,j=1}^{l+u} \|f(\boldsymbol{x_i}) - f(\boldsymbol{x_j})\|^2 \boldsymbol{W_{m,ij}} \\ &= \boldsymbol{f}^T \boldsymbol{L_m} \boldsymbol{f}, \end{aligned} \tag{8}$$

where $\boldsymbol{L_m}$ is the Laplacian matrix of $\boldsymbol{G_m}$. Likewise, the measure of characterizing the inter-class separability from the inter-class graph can be defined as follows:

$$\begin{aligned} \boldsymbol{S_c} &= \frac{1}{2} \sum_{i,j=1}^{l+u} \|f(\boldsymbol{x_i}) - f(\boldsymbol{x_j})\|^2 \boldsymbol{W_{c,ij}} \\ &= \boldsymbol{f}^T \boldsymbol{L_c} \boldsymbol{f}, \end{aligned} \tag{9}$$

where $\boldsymbol{L_c}$ is the Laplacian matrix of $\boldsymbol{G_m}$.

Here, a small $\boldsymbol{S_m}$ implies that every class has a small scatter. Meanwhile, a large $\boldsymbol{S_c}$ implies that different classes scatter well. Thus, we can introduce these regularization terms to our SSL framework.

Download English Version:

https://daneshyari.com/en/article/407716

Download Persian Version:

https://daneshyari.com/article/407716

Daneshyari.com