Pattern Recognition 44 (2011) 1777–1784

Contents lists available at ScienceDirect

Pattern Recognition

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

Sparse regularization for semi-supervised classification

Mingyu Fan [a], Nannan Gu [b], Hong Qiao [b], Bo Zhang [a,*]

[a] LSEC and Institute of Applied Mathematics, AMSS, Chinese Academy of Sciences, Beijing 100190, China
[b] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 21 June 2010
Received in revised form
29 September 2010
Accepted 15 February 2011
Available online 22 February 2011

Keywords:
Regularization theory
Semi-supervised learning
Regularized least square classification
Dimensionality reduction

ABSTRACT

Manifold regularization (MR) is a promising regularization framework for semi-supervised learning, which introduces an additional penalty term to regularize the smoothness of functions on data manifolds and has been shown very effective in exploiting the underlying geometric structure of data for classification. It has been shown that the performance of the MR algorithms depends highly on the design of the additional penalty term on manifolds. In this paper, we propose a new approach to define the penalty term on manifolds by the sparse representations instead of the adjacency graphs of data. The process to build this novel penalty term has two steps. First, the best sparse linear reconstruction coefficients for each data point are computed by the l^1-norm minimization. Secondly, the learner is subject to a cost function which aims to preserve the sparse coefficients. The cost function is utilized as the new penalty term for regularization algorithms. Compared with previous semi-supervised learning algorithms, the new penalty term needs less input parameters and has strong discriminative power for classification. The least square classifier using our novel penalty term is proposed in this paper, which is called the Sparse Regularized Least Square Classification (S-RLSC) algorithm. Experiments on real-world data sets show that our algorithm is very effective.

1. Introduction

Regularization theory was originally introduced to solve ill-posed inverse problems [18]. In the past decades, regularization has shown great power and been applied in many areas of machine learning, such as regression, clustering, classification and model selection [10]. Many state-of-art machine learning algorithms, including Support Vector Machines (SVMs) [19], Regularized Neural Networks (RNNs) [14] and Regularized Least Square Classifier (RLSC) [17], can be derived from the regularization framework.

Recently, in [4], Belkin et al. proposed a general Manifold Regularization (MR) framework for a full range of learning problems from unsupervised, semi-supervised to supervised. The framework was developed in the setting of Reproducing Kernel Hilbert Spaces (RKHS), and a new Representor theorem was obtained in this setting for the regularization framework. In contrast to the traditional regularization theory, which concentrates on the complexity of functions in the functional space, the MR framework supplements an additional penalty term to the traditional regularization based on the assumption that data lie on an intrinsic low-dimensional manifold. The additional penalty term is used to measure the smoothness of functions on data manifolds, which will be referred to as the (penalty) term on manifolds for short. Such a term can improve the performance of the obtained learner by exploiting the intrinsic structure of data. The MR algorithms, including the Laplacian Regularized Least Square Classification (LapRLSC) and the Laplacian SVM (LapSVM) methods [4], have been shown especially useful and efficient in semi-supervised learning problems when both labeled examples and unlabeled examples are available for learning.

Many semi-supervised learning methods can be unified in the MR framework. The Discriminatively Regularized Least Square Classification (DRLSC) method builds the penalty term on manifolds by integrating both discriminative and geometrical information in each local region [23]. Although the method is proposed as a supervised learning method, it can be applied to semi-supervised classification problems. The MR framework can also unify many of the graph-based semi-supervised learning algorithms by ignoring the complexity of functions, which only have the penalty term on manifolds in the framework. Zhu et al. proposed a semi-supervised learning method called the Gaussian fields and harmonic functions (GFHF) method, based on a Gaussian random field model [24]. Wang and Zhang proposed a semi-supervised learning algorithm by using the local linear reconstruction coefficients, which is similar to the GFHF method [20].

Despite the success of these semi-supervised classification methods, there are still some issues that have not yet been properly addressed. In particular,

(1) *Neighbors selection.* Many graph-based methods, including the MR framework, define the adjacency graphs by using a fixed neighborhood size for all the data points. However, a fixed

* Corresponding author. Tel.: +86 10 6265 1358.
E-mail addresses: fanminyu@amss.ac.cn (M. Fan), gunannan@gmail.com (N. Gu), hong.qiao@ia.ac.cn (H. Qiao), b.zhang@amt.ac.cn (B. Zhang).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.02.013

neighborhood size causes the difficulty of parameter selection and cannot be adaptive to uneven data.

(2) *Manifold assumption.* Many graph-based methods, including the MR framework, assume that high-dimensional data distribute on a low-dimensional manifold. However, for many types of data, we lack convincing evidence for the manifold structure.

(3) *Explicit classifier for new points.* Some graph-based methods do not have an explicit multi-class classifier for novel examples, which limits their application in on-line decision making tasks.

To address the above issues, we propose the sparse regularization (SR) approach for semi-supervised learning. A novel penalty term is defined using the sparse representation [22] of the data. With the novel penalty term, the approach can derive classifiers in the MR framework. Therefore, the proposed SR approach not only inherits the advantages of fewer parameters and highly discriminative ability from the sparse representation, but also has a natural out-of-sample extension for novel examples, which is inherited from the MR framework. Experiments on real-world data sets demonstrate the effectiveness and highly discriminative ability of our approach.

The rest of this paper is organized as follows. Some previous works are introduced in Section 2. The proposed SR approach and the derived Sparse Regularized Least Square Classification (S-RLSC) algorithm are presented in Section 3. Then in Section 4, experiments on benchmark real-world data sets are reported. Finally, we conclude this paper in Section 5.

2. Previous works

In a general semi-supervised classification problem, the training data set is represented as $\{(x_i,z_i),x_{l+j},i=1,\ldots,l,j=1,\ldots,u\}$, where l is the number of labeled data points, u is the number of unlabeled data points, $x_i\in\mathbb{R}^N$ is a data point and $z_i\in\{-1,1\}$ is the class label of x_i.

2.1. Regularization on explicit functions

Belkin et al. [4] proposed the MR framework based on the theory of RKHS. Assuming that f is a real-valued function in the RKHS $\mathcal{H}_K$, the MR framework can be expressed in the form

$$f^*=\underset{f\in\mathcal{H}_K}{\operatorname{argmin}}\left\{\frac{1}{l}\sum_{i=1}^{l}V(x_i,z_i,f)+\gamma_A\|f\|_K^2+\gamma_I\|f\|_I^2\right\}, \tag{1}$$

where V is some loss function, $\|f\|_K^2$ is the norm of the function in $\mathcal{H}_K$ which controls the complexity of the classifier and $\|f\|_I^2$ is the penalty term to regularize the smoothness of the function on manifolds. If

$$\|f\|_I^2=\frac{1}{(l+u)^2}\sum_{i,j=1}^{l+u}(f(x_i)-f(x_j))^2w_{ij}, \tag{2}$$

where w_{ij} are edge weights in the data adjacency graph, then it follows by the Representer Theorem [4, Theorem 2] that the solution of the optimization problem (1) admits the representation

$$f^*(x)=\sum_{i=1}^{u+l}\alpha_i k(x_i,x) \tag{3}$$

in terms of the labeled and unlabeled samples, where $k(\cdot,\cdot)$ is some Mercer kernel function associated with the RKHS $\mathcal{H}_K$. For different choices of loss function V and $\|f\|_I^2$, different MR algorithms can be derived from the MR framework (1). For example, if the loss function V is defined to the square loss function

$$V(x,z,f)=(z-f(x))^2$$

and $\|f\|_I^2$ is defined by (2), then the Laplacian Regularized Least Square Classifier (LapRLSC) can be obtained; if V is chosen as the hinge loss function

$$V(x_i,z_i,f)=\begin{cases}1-z_if(x_i) & \text{if } z_if(x_i)>0\\ 0 & \text{otherwise}\end{cases}$$

and $\|f\|_I^2$ is again given as in (2), then the Laplacian Regularized Support Vector Machines (LapSVMs) can be obtained (see [4]). By using the square loss function as the loss function V and by making the best use of the underlying discriminative and geometrical information of the data manifold to define the penalty term $\|f\|_I^2$, a new MR algorithm called the DRLSC algorithm was obtained in [23] from the MR framework (1). Although the DRLSC algorithm was proposed as a supervised learning algorithm, it is similar with the LapRLSC algorithm and can be used as a semi-supervised learning algorithm.

2.2. Regularization on implicit functions

If the parameter γ_A is set to be zero, then the second term of the MR framework (1) that controls the complexity of the classifier vanishes. As a result, the feasible function f in (1) is not restricted to being in the RKHS $\mathcal{H}_K$. In fact, the feasible function f can be any function; in particular, it can be required to be an unknown or implicit function satisfying that $z_i=f(x_i)$ for $i=1,\ldots,l$, so the error part $(1/l)\sum_{i=1}^{l}V(x_i,z_i,f)$ vanishes. Thus, the MR framework has only the penalty term $\|f\|_I^2$ on manifolds, where f is an unknown or implicit function satisfying that $z_i=f(x_i)$ for $i=1,\ldots,l$. For unlabeled data points x_{l+i} $(i=1,\ldots,u)$ define implicitly $z_{l+i}=f(x_{l+i})$ for $i=1,\ldots,u$, which are unknown and regarded as the labels of the unlabeled data points x_{l+i} (or values of the function f at x_{l+i}) $(i=1,\ldots,u)$. If we further define $\|f\|_I^2=(1/2)\sum_{i,j}^{l+u}w_{ij}(z_i-z_j)^2$, where w_{ij} are edge weights in the data adjacency graph as defined in Subsection 2.1, then the labels z_{l+i} of unlabeled data points can be computed by minimizing $\|f\|_I^2$. Therefore, the minimization of the penalty term on manifolds can also give new semi-supervised learning algorithms.

Belkin and Niyogi proposed a manifold learning based classifier [3], which is built by the eigenvectors of the Laplacian matrix. Zhu et al. introduced the GFHF method based on a random field model [24]. The GFHF method is defined on a weighted graph superimposed on the whole data set, which comprises both labeled and unlabeled data points. The pairwise similarities between the data points are defined as

$$w_{ij}=\exp\left(-\sum_{k=1}^{N}\frac{(x_{ik}-x_{jk})^2}{\sigma_k^2}\right),$$

where x_{ik} is the k-th component of the data point x_i and σ_k is the length-scale hyper-parameter for the k-th component.

Let $W=(w_{ij})$ be the $(l+u)\times(l+u)$ similarity matrix, let D be the diagonal matrix of order $l+u$ with $D_{ii}=\sum_{j=1}^{l+u}w_{ij}$ and let $L=D-W$. Then the GFHF method minimizes the quadratic function

$$E(Z)=\frac{1}{2}\sum_{i,j}w_{ij}(z_i-z_j)^2=Z^TLZ, \tag{4}$$

where $Z=(z_1,\ldots,z_l,z_{l+1},\ldots,z_{l+u})^T$ with $z_{l+i}=f(x_{l+i})$, $i=1,\ldots,u$. The similarity matrix W (and also the diagonal matrix D) can be split into four blocks:

$$W=\begin{pmatrix}W_{ll} & W_{lu}\\ W_{ul} & W_{uu}.\end{pmatrix}$$

Assume that $Z=(Z_l^TZ_u^T)^T$, where $Z_l=(z_1,\ldots,z_l)^T$ and $Z_u=(z_{l+1},\ldots,z_{l+u})^T$. Suppose Z minimizes the function in (4). Then