Neurocomputing 139 (2014) 345–356

Contents lists available at ScienceDirect

Neurocomputing

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

A kernel-based sparsity preserving method for semi-supervised classification

Nannan Gu [a], Di Wang [b,*], Mingyu Fan [b], Deyu Meng [c]

[a] School of Statistics, Capital University of Economics and Business, Beijing 100070, China
[b] Institute of Intelligent Systems and Decision, Wenzhou University, Wenzhou 325000, China
[c] Institute for Information and System Sciences and Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 7 January 2013
Received in revised form
12 January 2014
Accepted 7 February 2014
Communicated by Steven Hoi
Available online 8 April 2014

Keywords:
Sparse representation
Feature extraction
Manifold regularization
Semi-supervised learning
Semi-supervised classification

ABSTRACT

In this paper, we propose an effective approach to semi-supervised classification through kernel-based sparse representation. The new method computes the sparse representation of data in the feature space, and then the learner is subject to a cost function which aims to preserve the sparse representing coefficients. By mapping the data into the feature space, the so-called "l^2-norm problem" that may be encountered when directly applying sparse representations to non-image data classification tasks will be naturally alleviated, and meanwhile, the label of a data point can be reconstructed more precisely by the labels of other data points using the sparse representing coefficients. Inherited from sparse representation, our method can adaptively establish the relationship between data points, and has high discriminative ability. Furthermore, the new method has a natural multi-class explicit expression for new samples. Experimental results on several benchmark data sets are provided to show the effectiveness of our method.

1. Introduction

Semi-supervised classification (SSC), which aims to utilize both the labeled and unlabeled data simultaneously to train classifiers, has emerged in recent years since usually labeled samples are scarce and time-consuming to obtain, while unlabeled data are abundant and relatively easier to get. Moreover, under certain assumptions, it has been shown that the information conveyed by the marginal distribution of the unlabeled samples can help to boost the classification performance. Since SSC requires less human effort and gives higher accuracy, it has attracted considerable attention in the fields of data mining and machine learning. So far, many SSC methods have been proposed, such as the Expectation-Maximization algorithm for semi-supervised generative mixture models [1,2], self-training [3,4], co-training [5], and transductive support vector machines [6,7].

Among the various kinds of SSC approaches, the graph-based approaches have become one of the hottest research areas. They first model the whole data set as a graph and then perform classification on the graph based on certain assumptions such as the cluster assumption [8] which states that points are likely to belong to the same class if they locate in the same cluster. Some representative methods include the Gaussian random field (GRF) [9,10], the Local and Global Consistency (LGC) [11], and the Laplacian Regularized Least Square Classification (LapRLSC) [12]. Many graph-based SSC methods can be unified into the Manifold Regularization (MR) framework [12], which consists of a fitting term for the labeled points, a regularization term to control the complexity of the classifier, and another regularization term to control the smoothness of the classifier with respect to the geometric distribution of data.

Though graph-based SSC methods have been successfully applied in many fields [13], there are still certain issues that have not been properly solved. For example, the selection of neighborhood size for the adjacency graph is a difficult parameter selection problem, the manifold structure assumed by many graph-based methods is short of convincing evidence, and the explicit multi-class classifiers cannot be obtained for many graph-based methods.

To address the above issues, we propose to utilize the power of sparse representation of data. Sparse representation of data refers to the reconstruction of a data point by a linear combination of a small number of elementary points from a dictionary [14,15]. It has the advantage of high discriminative power and adaptively establishing the relationship between data points [16]. Recently, there have emerged several classification methods based on sparse

* Corresponding author.
E-mail addresses: nngu@cueb.edu.cn (N. Gu), wangdi@wzu.edu.cn (D. Wang), fanmingyu@wzu.edu.cn (M. Fan), dymeng@mail.xjtu.edu.cn (D. Meng).

http://dx.doi.org/10.1016/j.neucom.2014.02.022
0925-2312/

representations [17–19]. These methods have shown great discriminative ability for image data sets; however, the following two issues still need to be settled:

(1) *The l^2-norm problem*: For general natural data sets where different points may have different l^2-norms, the sparse representation of one data point may be inclined to select the points with larger l^2-norms other than the ones in the same class (the reason will be given in Section 3.1). Moreover, normalizing all points to have a unit l^2-norm may change the structure of the data set.
(2) *The imprecise reconstruction of labels by sparse representations*: Sparsity-based methods [18,19] are usually based on the fact that the label of a data point can be reconstructed by the labels of other points using the coefficients of sparse representation. However, the classification function is usually nonlinear, so the reconstruction is not precise, which may in return decrease the classification accuracy of the obtained classifier.

In view of this, we propose the Kernel-based Sparse Regularization (KSR) approach to semi-supervised learning. First, the data points are projected into the kernel space by the kernel trick, and then the sparse representation of each projected point is computed in the kernel space. Finally, inspired by the MR framework, the multi-class classifier is constructed based on the discriminative ability of the obtained sparse representation. In this way, the projected data points in the kernel space naturally have a unit l^2-norm, which can overcome the l^2-norm problem described above. Meanwhile, the classification function is linear in the kernel space so the sparse representing coefficients can be preserved more precisely. The proposed KSR approach not only has an explicit formulation for multi-classification problem and can be easily utilized in any multi-class (including binary-class) cases, but also inherits the advantage of high discriminative ability and adaptively establishing the relationship between data points from sparse representation. Experiments on real-world data sets demonstrate the effectiveness and high discriminative ability of our approach.

The rest of the paper is organized as follows. Some previous work is introduced in Section 2. The proposed KSR approach and the derived Kernel-based Sparse Regularized Least Square Classification (KSR-LSC) algorithm are presented in Section 3. In Section 4, experiments on benchmark real-world data sets are reported. Final conclusions are given in Section 5.

2. Previous work

In this paper, we represent the training data set of the semi-supervised classification problem as $\{(x_i, z_i),\ x_{l+j}, i=1,\ldots,l,\ j=1,\ldots,u\}$, where l and u are the number of labeled and unlabeled data points, respectively, $x_i \in \mathbb{R}^d$ is a data point, $z_i \in \{1,\ldots,C\}$ is the class label of x_i and C is the total number of classes. Throughout this paper, the data points and the corresponding label vectors are in the form of column vectors. All vectors are indicated by lowercase letters and all matrices are denoted by normal capital letters.

2.1. The MR framework

The MR framework [12] was proposed for learning based on the theory of Reproducing Kernel Hilbert Space (RKHS). It can be expressed in the form

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

where f is the desired classification function, V is some loss function, $\|f\|_K^2$ is the square of the norm of f in the RKHS $\mathcal{H}_K$ to regularize the complexity of the classifier and $\|f\|_I^2$ is another regularization term to control the smoothness of the classifier on the data manifold. If we define

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

where w_{ij} is the edge weight in the data adjacency graph, then the solution of (1) admits the representation [12]

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

where $k(\cdot,\cdot)$ is some Mercer kernel function associated with the RKHS $\mathcal{H}_K$.

Many SSC algorithms can be unified into the MR framework. However, there are still some issues to be addressed. For example, usually a fixed neighborhood size is adopted to define the adjacency graphs, which brings the difficulty of parameter selection and cannot be adaptive to uneven data. Besides, the methods are often based on the assumption that high-dimensional data distribute on a low-dimensional manifold. However, in many cases, this assumption may not be true, so several classification methods have been proposed based on sparse representation. These methods can adaptively establish the relationship between data points and have high discriminative ability.

2.2. Sparse representation based classification methods

Sparse representation of data refers to constructing each data point with a linear combination of a small number of elementary points from a dictionary. For image data, if sufficient training samples are available for each class, it will be possible to represent one sample as a linear combination of just those samples from the same class. This representation is naturally sparse, involving only a small fraction of the overall training samples, i.e., the samples from the same class. So the sparse representation of data is discriminative and can be utilized to train the classifier.

Based on the discriminative ability of sparse representation of data, Wright et al. proposed a general classification algorithm named SRC for (image-based) object recognition [17], where each test sample is represented by an over-complete dictionary whose base elements are the training samples themselves. Specifically, the sparse representation of data is obtained by the l^1-norm minimization problem [20–22]

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^N} \|\alpha\|_1 \qquad \text{subject to} \quad x^* = X\alpha, \tag{2}$$

where x^* is a test sample, α^* is the desired sparse representation of x^*, $X = [X_1, \ldots, X_C] \in \mathbb{R}^{d \times N}$ is the training data matrix of C classes and $\|\cdot\|_1$ is the l^1-norm of vectors. For $\alpha^* \in \mathbb{R}^N$, denote by $\delta_i(\alpha^*) \in \mathbb{R}^N$ the vector whose only nonzero entries are the entries in α^* that are associated with class i. Then they classify x^* as

$$\text{identity}(x^*) = \arg\min_{i \in \{1,\ldots,C\}} \|x^* - X\delta_i(\alpha^*)\|_2.$$

Fan et al. proposed an approach (named S-RLSC) to semi-supervised learning based on sparse representation [18]. In [18], the best sparse linear reconstruction coefficient vector $\alpha_i \in \mathbb{R}^{l+u}$ is first computed for each data point x_i $(i=1,\ldots,l+u)$ via the l^1-norm minimization (2). The sparse representation is naturally discriminative, and it is meaningful to require that the classification function f on the data points keeps the sparse representing coefficients. Hence, the label of a data point can be reconstructed by the labels of other data points using the sparse representing