Neurocomputing 131 (2014) 124–131

Contents lists available at ScienceDirect

Neurocomputing

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

Semi-supervised learning via sparse model

Yu Wang [a,b,*], Sheng Tang [a], Yan-Tao Zheng [c], Yong-Dong Zhang [a,*], Jin-Tao Li [a]

[a] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[b] Graduate University of Chinese Academy of Sciences, Beijing 100190, China
[c] Institute for Infocomm Research, A*STAR, Singapore 138632, Singapore

ARTICLE INFO

Article history:
Received 19 November 2012
Received in revised form
24 October 2013
Accepted 26 October 2013
Communicated by Marco Loog
Available online 1 December 2013

Keywords:
Manifold structure
Cluster structure
Semi-supervised learning
Sparse model

ABSTRACT

Graph-based Semi-Supervised Learning (SSL) methods are the widely used SSL methods due to their high accuracy. They can well meet the manifold assumption with high computational cost, but don't meet the cluster assumption. In this paper, we propose a Semi-supervised learning via SPArse (SSPA) model. Since SSPA uses sparse matrix multiplication to depict the adjacency relations among samples, SSPA can approximate low dimensional manifold structure of samples with lower computational complexity than these graph-based SSL methods. Each column of this sparse matrix corresponds to one sparse representation of a sample. The rational is that the inner product of sparse representations can also be sparse under certain constraint. Since the dictionary in the SSPA model can depict the distribution of the entire samples, the sparse representation of a sample encodes its spatial location information. Therefore, in the SSPA model the manifold structure of samples is computed via their locations in the intrinsic geometry of the distribution instead of their feature vectors. In order to meet the cluster assumption, we propose an structured dictionary learning algorithm to explicitly reveal the cluster structure of the dictionary. We develop the SSPA algorithms with the structured dictionary besides non-structured one, and experiments show that our methods are efficient and outperform state-of-the-art graph-based SSL methods.

1. Introduction

Semi-supervised learning (SSL) uses a large number of unlabeled samples together with a small number of labeled samples to build a better learner. It has attracted great research interest in both theory and practice in the past decade [1]. There exist various approaches in SSL to exploit unlabeled samples, showing incremental performance improvements [2]. Among them, graph-based methods, also known as manifold methods [3,4], are widely used due to their good performances [3,5,6].

Through establishing the graph structure to represent adjacency relations among samples, the graph-based SSL methods can approximate the intrinsically low-dimensional manifold structure well [7]. Existing ways for graph construction include K-nearest-neighbor (KNN) method and ε-ball based method [5,6]. In general, they construct the graph structure in two steps: (1) choosing adjacency samples needed to be connected, and (2) determining the edge weights [6]. l_1 graph, recently proposed by [4], is constructed at one run by solving each sample's sparse representation based on the rest of the entire sample set. It can surpass KNN method and ε-ball based method due to its characteristics of the high discriminative power, sparsity, and adaptive neighborhood [8]. However, these graph-based SSL methods have common drawbacks: high computational cost always comes together with the graph construction. When the adjacency relations among samples are computed to construct the graph structure, the complexity is $O(mn^2)$ [3], where n is the number of data samples and m is the dimensionality of the sample feature.

The graph-based SSL methods do not meet the cluster assumption. The cluster assumption states that if points are located in the same cluster, they are likely to belong to the same class. The success of SSL relies highly on certain semi-supervised assumptions (SSA) on the samples' distribution, which includes manifold and cluster assumptions [1]. The manifold assumption indicates that the high dimensional data lie on a low-dimensional manifold. Graph-based methods have not explored the cluster information of input distribution, so they may connect data points of different classes in cluster boundaries by the Euclidean distances of their feature vectors.

In order to overcome the above shortcomings of the graph-based SSL methods, we propose a new Semi-supervised learning method based on SPArse model, named SSPA. With the SSPA model, following advantages are achieved: Firstly adjacency relations among samples can be achieved by sparse matrix multiplication with a complexity of

* Corresponding author at: Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China.
Tel.: +86 010 6260 0666; fax: +86 010 6260 0488 810.
E-mail addresses: wangyu@ict.ac.cn (Y. Wang), ts@ict.ac.cn (S. Tang), yantaozheng@gmail.com (Y.-T. Zheng), zhyd@ict.ac.cn (Y.-D. Zhang), jtli@ict.ac.cn (J.-T. Li).

0925-2312/$ - see front matter
http://dx.doi.org/10.1016/j.neucom.2013.10.033

$O(sn^2)$ [9], where s is the average number of nonzero elements in each sparse representation and far less than the feature dimensionality m, so $O(sn^2) \ll O(mn^2)$. Secondly, we explore the fact that the dictionary in SSPA model has the same cluster information as the training data samples, and thus we propose a structured dictionary learning algorithm to learn a structured dictionary which has explicit cluster information. With the structured dictionary the structured sparse model can be used in SSPA to gain better accuracy.

The main contribution of the work is that we propose a new SSPA model, with two kinds of dictionaries: non-structured [10] and structured. The SSPA model has a lower computational complexity than the graph-based SSL methods. With the structured dictionary the SSPA model can utilize the cluster information to improve the classification accuracy. The experimental results demonstrate that the proposed methods are efficient and outperform state-of-the-art graph-based SSL methods.

2. Related work

This work is closely related to the graph-based SSL methods and sparse model.

2.1. Graph-based SSL methods

Graph-based SSL methods make good use of manifold structure to explore the geometric structure behind labeled and unlabeled data. In general, firstly a graph is formed over the labeled and unlabeled points, and then the label information can be propagated over the graph. Zhu and Ghahramani proposed a label propagation algorithm [5] through the graph representation. Szummer and Jaakkola presented a closely related approach with [5] which used random walks through the graph to assign labels [11]. Deng and He used the graph structure as a prior to guide semi-supervised Discriminant Analysis [12]. Bekin and Niyopi introduced a Manifold Regularization (MR) framework which, in contrast to purely graph-based SSL methods, had the capability of out of samples extension [6]. Ni et al. proposed a criteria to construct a graph with the pursued optimal feature representation for SSL [13]. Liu et al. used linear combinations of a small number of anchor points to construct a large graph for SSL [14]. Though the graph structure is widely used, it does not always bring positive effects for learning and sometime it may hurt the learning results. Some examples are illustrated in our experiments.

The traditional graph construction methods do not explore the cluster information of input distribution. However, cluster structures can benefit the graph construction. Specifically cluster information can help to remove the adjacency links among different clusters. By exploring the cluster information the proposed methods have the ability to depict the whole structures of input space (i.e. multiple clusters and manifolds), and compare samples with their location information.

2.2. Sparse model

Recent researches have shown that sparsity can help to improve the performances of various machine learning problems, such as the classification problem [15] and the feature extraction problem [16,17]. Existing sparse model methods can be generally classified into the following two categories:

(1) The entire database is used as the dictionary, and each data point can be represented by a linear combination of the rest data points. Wright et al. [18] demonstrated that sparse representation can be used for direct face recognition. Elhamifar exploited the block structure of the training database with structured sparse representation [19]. Sparse representation on the training database was used to cluster data by Elhamifar and Vidal [20].

(2) Dictionary is learnt from data. Dictionaries, which were learned from different class databases, had been used for clustering problem [21,22]. Raina et al. proposed a method to learn a common dictionary from a few databases to extract new features for transfer learning task [23]. Mairal and Bach [24] proposed an online optimization algorithm based on stochastic approximation to learn the dictionary from large-scale databases.

The l_1 graph proposed by [4] belongs to the first category. It is solved by l_1 norm optimization on the whole training samples, and has been used for SSL problem [25]. However, using the whole training database as a large redundant and coherent dictionary makes l_1 graph unstable and expensive [21]. Thus, in our proposed SSPA model we learn the dictionary to provide more stable and effective l_1 norm optimization. Specifically, we have implemented SSPA algorithms with two kinds of dictionaries: (1) non-structured dictionary learned by [10] is used to ensure that the distribution of dictionary coincides with that of the whole training set; (2) structured dictionary learned by the proposed algorithm in Section 3.3 is used to meet the cluster assumption.

3. SSPA model

This section elaborates on the formulation of the SSPA model, and theoretically proves that the proposed SSPA model is capable of depicting the cluster and manifold structures of data samples.

3.1. Overview

With a set of l labeled samples $(x_i, y_i)_{i=1}^{l}$ and a set of u unlabeled samples $(x_j)_{j=l+1}^{l+u}$, the SSPA model is given by

$$f^* = \arg\min_{f \in H_k} \frac{1}{l} \sum_{i=1}^{l} V(z(x_i), y_i, f) + r_A \|f\|_k^2 + \frac{r_I}{(u+l)^2} \sum_{i,j=1}^{l+u} (f(z(x_i)) - f(z(x_j)))^2 M_{ij}, \quad (1)$$

where M is the adjacency matrix, V is a standard loss function on labeled samples and $z(x_i)$ is an optimization representation of feature vectors. We apply the representation introduced by Balcan et al. [26] in this work. Our SSPA model aims to measure the similarity between two samples through a special l_1 norm optimization problem, which is the following sparse model proposed by [10]

$$g(x_j, D) = \arg\min \|\alpha_j\|_1, \quad \text{s.t.} \quad D\alpha_j = x_j, \ D \in X, \quad (2)$$

where $D \in R^{m \times k} (m < k)$ is the over-complete dictionary consisting of k bases, $\alpha \in R^k$ is the sparse representation of feature vector $x \in R^m$, and data matrix $X = \{x_1, \ldots, x_n\}$.

Those atoms of D learned from samples through traditional dictionary learning algorithms [24,27] are easy to be excessively coherent, and can not coincide with the original data items. Thus, we restrict those atoms to be the subset of the database as in Eq. (2) ($D \in X$). Specifically, those atoms of D are on behalf of the databases, and the dictionary D in Eq. (2) can be learned by our former work [10] to depict the cluster structure of data samples. For arbitrary pair of samples x_p, x_t, their similarity weight is

$$M(x_p, x_t) = (\alpha_p \alpha_t^T) = (g(x_p, D)^T g(x_t, D)) \quad (3)$$

Manifold structure of the training samples can be achieved by Eq. (3), due to the characteristics of SSPA model as detailed in next subsection. The adjacency matrix M is computed through a sparse