Neural Networks 51 (2014) 9–16

Contents lists available at ScienceDirect

Neural Networks

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

Feature selection and multi-kernel learning for sparse representation on a manifold

Jim Jing-Yan Wang [a,b], Halima Bensmail [c], Xin Gao [a,d,*]

[a] Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

[b] Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

[c] Qatar Computing Research Institute, Doha 5825, Qatar

[d] Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

ARTICLE INFO

Article history:
Received 27 January 2013
Received in revised form 23 June 2013
Accepted 13 November 2013

Keywords:
Data representation
Sparse coding
Manifold
Feature selection
Multiple kernel learning

ABSTRACT

Sparse representation has been widely studied as a part-based data representation method and applied in many scientific and engineering fields, such as bioinformatics and medical imaging. It seeks to represent a data sample as a sparse linear combination of some basic items in a dictionary. Gao et al. (2013) recently proposed Laplacian sparse coding by regularizing the sparse codes with an affinity graph. However, due to the noisy features and nonlinear distribution of the data samples, the affinity graph constructed directly from the original feature space is not necessarily a reliable reflection of the intrinsic manifold of the data samples. To overcome this problem, we integrate feature selection and multiple kernel learning into the sparse coding on the manifold. To this end, unified objectives are defined for feature selection, multiple kernel learning, sparse coding, and graph regularization. By optimizing the objective functions iteratively, we develop novel data representation algorithms with feature selection and multiple kernel learning respectively. Experimental results on two challenging tasks, N-linked glycosylation prediction and mammogram retrieval, demonstrate that the proposed algorithms outperform the traditional sparse coding methods.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

In recent years, sparse representation as a part-based representation method has attracted much attention from both the academic and industrial communities (Wang, Bensmail, Yao, & Gao, 2013). Sparse representation methods assume that a data sample can be represented as a sparse linear combination of some basic elements in a dictionary. The resulting optimization problem penalizes the l_1-norm of the linear combination coefficients (Wright, Yang, Ganesh, Sastry, & Ma, 2009). In Wright et al. (2009), Wright et al. proposed the use of sparse representation for the robust face recognition problem by representing a face image as the sparse reconstruction of the training face images in the database. Sparse coding (Sc) was further proposed by Lee, Battle, Raina, and Ng (2007), in which they not only learned the sparse reconstruction coefficients but also the dictionary containing the basic elements. Moreover, non-negative matrix factorization (NMF) (Wang, Almasri, & Gao, 2012) was improved to sparse NMF via alternating non-negativity-constrained least squares by Kim and Park (2007).

By using the sparse constraints, both sparse coding and sparse NMF can learn a good part-based representation. However, they perform the learning in the Euclidean space and fail to discover the intrinsic manifold structure of the data space, which is essential to real-world applications (Cai, He, Han, & Huang, 2011). Because of the sparsity of the combination coefficients and the overcompleteness of the dictionary, the similar data samples may be encoded as totally different sparse codes by sparse coding or sparse NMF. To address this problem, manifold constraints were imposed on the sparse representation method, by explicitly taking into account the local manifold structure of the data. For example, Cai et al. (2011) proposed the graph-regularized non-negative matrix factorization (GNMF) for data representation by constructing an affinity graph to encode the geometrical information and by seeking a matrix factorization, that respected the graph structure. Similarly, Gao, Tsang, and Chia (2013) proposed the Laplacian sparse coding (LapSc) by incorporating a graph-based local similarity preserving term into the objective function of sparse coding, thus releasing the instability of the sparse codes.

Most manifold regularized representation methods employ the nearest neighbor graph structure to encode the manifold information (Cai et al., 2011; Gao et al., 2013). The graph is constructed from the original feature space of samples and then used to regularize the sparse representations, assuming that if two data samples are close in the original feature space, the representations of

* Corresponding author at: Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. Tel.: +966 2 808 0323.

E-mail addresses: jimjywang@gmail.com (J.J.-Y. Wang), xin.gao@kaust.edu.sa (X. Gao).

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.11.009

these two samples are also close to each other. Despite the success of such a common graph strategy for manifold regularization, two major problems have not yet been properly addressed:

1. Some features from the original feature space are *noisy features*, which are irrelevant to the tasks at hand. Graphs constructed using these features cannot reflect the intrinsic manifold structure of the data samples. One may first perform the feature selection and then use the selected features to construct the graph for sparse representation regularization. However, the sparse representation on the manifold cannot adjust itself according to the feature selection results. It also cannot capture the intrinsic relations among the features, the sparse representation and the manifold.
2. The original data may lie on a *nonlinear distribution*, although the neighbors are found from the linear distance between the data samples, like the Euclidean distance (Courrieu, 2005), thus making the nearest neighbor graph not necessarily an accurate representation of the intrinsic manifold structure. A kernel trick was recently proposed to handle this problem by mapping the data to a high-dimensional, nonlinear Hilbert space (Alzate & Suykens, 2012). However, the selection of kernels and parameters remains a difficult problem. One possible way to solve this problem is to use cross-validation to select the optimal kernel, but it suffers from being time consuming and are easily overfitted.

To overcome the disadvantages mentioned above and inspired by Zeng and Cheung (2011), we investigate the intrinsic relation between the feature selection/multiple kernel learning and the sparse representation on a manifold. The contribution of this paper consists of two novel manifold-regularized sparse representation algorithms dealing with noisy features and nonlinearly distributed data respectively:

- To handle the noisy features, we propose a novel method that performs feature selection within the framework of the manifold-regularized sparse representation. We propose a novel unified objective function that takes into account the feature selection, sparse representation and manifold regularization simultaneously. In the sparse representation, the feature weights and the graph are updated alternately by optimizing the objective function, resulting in a novel Sparse Representation on a Manifold with Feature Selection algorithm—**SRM-FS**.
- To handle the nonlinear distribution data and the kernel and parameter selection, we propose a novel method that integrates the multiple kernel learning and the sparse representation regularized by the manifold. Given a pool of kernels with different model definitions and parameters, the final kernel is learned by the weighted linear combination of the kernels from this pool. The multiple kernel weights, the sparse representations and the affinity graph are learned by optimizing a unified objective function alternately, resulting in a novel Sparse Representation on a Manifold with Multiple Kernel Learning algorithm—**SRM-MKL**.

The remainder of this paper is organized as follows: We present the proposed sparse representation algorithm on a manifold with feature selection in Section 2, and then extend it to the sparse representation algorithm on a manifold with multiple kernel learning in Section 3. In Section 4, comparative experiments on two challenging tasks are conducted to show the performance of the proposed methods. Finally, conclusions are drawn in Section 5.

2. Feature selection for sparse representation on a manifold

Suppose we have n samples in the training dataset denoted as $\mathcal{X} = \{x_1, \ldots, x_n\}$, where $x_i = [x_{i1}, \ldots, x_{id}]^\top \in \mathbb{R}^d$ is the d-dimensional feature vector of the i-th sample. The task of feature selection is to scale the features with different weights and thus obtain a weighted feature space, parameterized by a nonnegative vector, $\lambda = [\lambda_1, \ldots, \lambda_d]^\top \in \mathbb{R}_+^d$, where λ_l is the scaling weight of the l-th feature, restricted by $\sum_{l=1}^d \lambda_l = 1$ (Gold, Holub, & Sollich, 2005). The feature vector of the i-th sample weighted by λ can be denoted as $x_i^\lambda = [\lambda_1 x_{i1}, \ldots, \lambda_d x_{id}]^\top = \mathrm{diag}(\lambda)x_i$, where $\mathrm{diag}(\lambda) \in \mathbb{R}_+^{d\times d}$ is a diagonal matrix with λ on the diagonal.

Given the training set, $\mathcal{X}$, the sparse representation aims to find a set of basic vectors, $\mathcal{U} = \{u_1, \ldots, u_m\} \in \mathbb{R}^d$, such that each training sample, x_i, can be represented as a sparse linear combination of those basic vectors in the dictionary, $\mathcal{U}$, as

$$x_i \approx \sum_{k=1}^m u_k v_{ik}, \tag{1}$$

where $v_{ik}, k = 1, \ldots, m$ are the linear combination coefficients (Wright et al., 2009), which are supposed to be as sparse as possible. By denoting $U = [u_1, \ldots, u_m] \in \mathbb{R}^{d\times m}$ as the basic matrix and $v_i = [v_{1i}, \ldots, v_{mi}] \in \mathbb{R}^m$ as the coefficient vector for the i-th sample, and also integrating the feature scaling weights in λ to both the sample vectors and basic vectors, (1) can be turned to $\mathrm{diag}(\lambda)x_i \approx \mathrm{diag}(\lambda)Uv_i$. Note that the feature selection has also been performed on the basic vectors of the dictionary. The coefficient vector, v_i, can be regarded as the new representation of the i-th sample in this new space with respect to the weighted basic vectors.

We further hope that the sparse representation respects the intrinsic manifold structure of the λ-weighted feature space. The manifold structure is represented by a nearest neighbor graph, G^λ, constructed from the training set. G^λ has n nodes, and each node represents a sample. The graph affinity matrix, $W^\lambda = [W_{ij}^\lambda] \in R^{n\times n}$, can be constructed from the n λ-weighted sample feature vector using a Gaussian kernel, $W_{ij}^\lambda = \exp(-\frac{\|x_i^\lambda - x_j^\lambda\|^2}{\sigma_{ij}})$ (Belanovic, Valcarcel Macua, & Zazo, 2012), if x_j^λ is among the nearest neighbors of x_i^λ, and 0 otherwise.

To obtain the optimal feature weights, λ, and the sparse representation $\{v_i\}$ for samples in $\mathcal{X}$ with the corresponding dictionary, U, simultaneously, we propose the following optimization problem:

$$\begin{aligned}\min_{U,\{v_i\},\lambda}\ & \sum_{i=1}^n \|\mathrm{diag}(\lambda)(x_i - Uv_i)\|^2 + \alpha\sum_{i=1}^n \|v_i\|_1 \\ & + \beta\sum_{i,j=1}^n \|v_i - v_j\|^2 W_{ij}^\lambda \\ s.t.\ & \sum_{l=1}^d \lambda_l = 1, \quad \lambda_l \geq 0.\end{aligned} \tag{2}$$

The objective function to be minimized above is composed of three terms and weighted by trade-off parameters, α and β, which could be set by cross-validation. The first term is to preserve the sample fidelities between the $\mathrm{diag}(\lambda)x_i$ and its approximation $\mathrm{diag}(\lambda)Uv_i$. The second l_1-norm-based term determines that the representation coefficient vectors, v_i, are sparse. The last term is used to embed the manifold information into the sparse representations. If two weighted sample feature vectors, x_i^λ and x_j^λ, are close in the intrinsic geometry of the data distribution, i.e, W_{ij}^λ is big, then v_i and v_j, the sparse representations of these two samples, are also close to each other. Moreover, the constraints $\sum_{l=1}^d \lambda_l = 1,\ \lambda_l \geq 0$ are applied to λ to prevent negative contributions by the features or shrinking of feature weights. This formula makes two main contributions:

- The first is to apply the feature weighting to both the original feature vectors and the basic vectors. This is implemented

Download English Version:

https://daneshyari.com/en/article/404000

Download Persian Version:

https://daneshyari.com/article/404000

Daneshyari.com