Neurocomputing 204 (2016) 135–141

Contents lists available at ScienceDirect

Neurocomputing

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

Manifold regularized multi-view feature selection for social image annotation

Yangxi Li [a], Xin Shi [b], Cuilan Du [a], Yang Liu [a,*], Yonggang Wen [c]

[a] National Computer network Emergency Response technical Team/Coordination Center of China (CNCERT/CC), China
[b] Key Laboratory of Machine Perception, Peking University, Beijing, China
[c] School of Computer Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 15 March 2015
Received in revised form
4 June 2015
Accepted 13 July 2015
Communicated by Dacheng Tao
Available online 16 April 2016

Keywords:
Feature selection
Multi-view
Manifold regularization
Image annotation

ABSTRACT

The features used in many social media analysis-based applications are usually of very high dimension. Feature selection offers several advantages in highly dimensional cases. Recently, multi-task feature selection has attracted much attention, and has been shown to often outperform the traditional single-task feature selection. Current multi-task feature selection methods are either supervised or unsupervised. In this paper, we address the semi-supervised multi-task feature selection problem. We firstly introduce manifold regularization in multi-task feature selection to utilize the limited number of labeled samples and the relatively large amount of unlabeled samples. However, the graph constructed in manifold regularization from a single feature representation (view) may be unreliable. We thus propose to construct the graph using the heterogeneous feature representations from multiple views. The proposed method is called manifold regularized multi-view feature selection (MRMVFS), which can exploit the label information, label relationship, data distribution, as well as correlation among different kinds of features simultaneously to boost the feature selection performance. All these information are integrated into a unified learning framework to estimate feature selection matrix, as well as the adaptive view weights. Experimental results on three real-world image datasets, NUS-WIDE, Flickr and Animal, demonstrate the effectiveness and superiority of the proposed MRMVFS over other state-of-the-art feature selection methods.

1. Introduction

Feature selection lies at the heart of many multimedia analysis-based applications, such as image search [24,25,27], automatic image annotation [10,12,13], image classification [23,26], etc. In these applications, where high dimensional features are usually utilized, a compact features selected from the original features is helpful to reduce the computational cost, save the storage space, and reduce the chance of over-fitting. Although traditional feature selection methods usually select features from a single task [3,7,16,28], more recently there has been a focus on joint feature selection across multiple related tasks [11,17,18,20,22]. This is because joint feature selection can exploit task relationship in order to establish the importance of features, and this approach has been empirically demonstrated to be superior to feature selection on each task separately [18].

Current multi-task feature selection methods are conducted either in a supervised [17,18] or unsupervised [11,20,22] manner, in terms of whether the label information is utilized to guide the selection of useful features. Supervised feature selection methods always require a large amount of labeled training data, and it may fail to identify the relevant features that are discriminative when the number of labeled samples is small. On the other hand, unsupervised feature selection methods are often unable to identify the discriminative features since the label information is ignored [21]. Since the cost of manually labeling of multi-view data is high, while large amount unlabeled multi-view data can easily be obtained, it is desirable to develop multi-view feature selection methods that are capable of exploiting both labeled and unlabeled data. This motivates us to introduce semi-supervised learning into the multi-view feature selection method. Therefore, we focus on the semi-supervised multi-task feature selection in this paper.

To make use of both the limited labeled data and the abundant unlabeled data, we employ the idea of manifold regularization [1] into multi-task feature selection. The performance of manifold regularization relies much on the constructed graph Laplacian. However, the graph constructed using the features from a single

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2015.07.151
0925-2312/

view may be unreliable. For example, the different labels cannot be properly characterized by a single feature representation in image classification [15]. We thus further extract different kinds of features from the image, and weightedly combine the graphs constructed using the different feature representations. The combined graph is able to approximate the underlying data distribution more accurately than the single graph.

The proposed method is called manifold regularized multi-view feature selection (MRMVFS). To the best of our knowledge, little progress has been made in semi-supervised multi-view feature selection. There are some recent works focused on unsupervised multi-view feature selection [5,20], and we mainly differ from them in that the label information is incorporated. On the other hand, compared to the works focused on semi-supervised feature selection [3,28], we mine the relationship information among multiple representations (views) and integrate it into feature selection jobs.

The proposed MRMVFS integrates four kinds of information, i.e., label information, label relationship, data distribution, as well as correlation among different views, to select the most representative feature components from the original multi-view features. In particular, a regression model is adopted to exploit the label information contained in the labeled samples. The $l_{2,1}$-norm penalty on the feature selection matrix is enforced to joint feature selection across multiple labels, and thus explore the label relationship. Meanwhile, the visual similarity graphs of different views are constructed and combined to model the geometric structure of the underlying data distribution. Besides, a set of non-negative view weights is learned to leverage the correlation among different views, and establish a reliable regularization term along the data manifold to smooth the prediction function. Finally, we integrate all these information into a unified learning framework. Based on this framework, we can simultaneously estimate feature selection matrix, as well as the adaptive view weights. In the experiments, we apply MRMVFS to automatic image annotation on three challenge image datasets, NUS-WIDE-OBJECT [4], Flickr [8] and Animal [9], and compare it with several state-of-the-art feature selection methods. Experimental results demonstrate the effectiveness and superiority of the proposed MRMVFS.

2. Manifold regularized multi-view feature selection

In this section, we elaborate the proposed manifold regularized multi-view feature selection (MRMVFS) method in detail.

2.1. Notations

We firstly introduce some important notations used in the rest of this paper. A matrix is represented by a capital letter, e.g., X. X_{ij} is the (i,j)th element of X, and $X_{i:}$ indicates the elements in the ith row of X. The bold lower case letter $\mathbf{x}$ indicates a vector and x indicates a scalar. Superscript indicates the view of data, e.g., $X^{(v)}$ is the vth view of data X. Subscript is used to denote if the data is labeled, for example, X_L is the labeled data, whereas X_U is the unlabeled data. $\|X\|_F$ denotes the matrix X's Frobenius norm. Specifically, for a matrix $X \in \mathbb{R}^{p\times q}$, its $l_{2,1}$-norm is defined as:

$$\|X\|_{2,1} = \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{q} X_{ij}^2} \tag{1}$$

2.2. Problem formulation

Given a set of l labeled samples $\mathcal{D}_L = \{(\mathbf{x}_i, y_i)_{i=1}^{l}\}$ and a relatively large set of u unlabeled samples $\mathcal{D}_U = \left\{(\mathbf{x}_i)_{i=l+1}^{l+u}\right\}$, we suppose each sample is represented by m different views, i.e., $\mathbf{x}_i = \left[\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(m)}\right]$. The view we refer to here is a certain kind of feature or modality. Then the feature matrix of the vth view can be represented as $X^{(v)} = [\mathbf{x}_1^{(v)}, \mathbf{x}_2^{(v)}, \ldots, \mathbf{x}_n^{(v)}] \in \mathbb{R}^{d_v\times n}$, and feature matrix of all the views is $X = \left[X^{(1)}, X^{(2)}, \ldots, X^{(m)}\right] \in \mathbb{R}^{d\times n}$, where $d = \sum_{v=1}^{m} d_v$ and d_v is the feature dimension of the vth view.

To select the compact and representative feature components from raw features, we propose to integrate four kinds of information, i.e., the label information contained in labeled data, label relationship, data distribution, as well as correlation among different views of both labeled and unlabeled data, into the learning framework.

We firstly introduce how to integrate label information of labeled data into MRMVFS. Given the labeled feature matrix $X_L = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l] \in \mathbb{R}^{d\times l}$ and the corresponding label matrix $Y_L = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_l]^T \in \mathbb{R}^{l\times c}$ with each $\mathbf{y}_i = [y_i^1, y_i^2, \ldots, y_i^c]^T$, we can learn the prediction functions $f^p(\mathbf{x}), p = 1, \ldots, c$ by minimizing the prediction error over the labeled data in training set

$$\min_{\{f^p\}} \sum_{p=1}^{c} \sum_{i=1}^{l} \mathcal{L}(f^p(\mathbf{x}_i), y_i^p), \tag{2}$$

where $\mathcal{L}$ is some pre-defined convex loss, $y_i^p = 1$ if the pth label is manually assigned to the ith sample, and -1 otherwise. We assume each $f^p(\mathbf{x})$ is a linear transformation with $f^p(\mathbf{x}_i) = (\mathbf{w}^p)^T\mathbf{x}_i$. Let $\mathbf{f}(\mathbf{x}) = [f^1(\mathbf{x}), f^2(\mathbf{x}), \ldots, f^c(\mathbf{x})]^T$ we have $\mathbf{f}(\mathbf{x}) = W^T\mathbf{x}$, where $W = [\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^c] \in \mathbb{R}^{d\times c}$ is the transformation matrix. To make it suitable for feature selection, a $l_{2,1}$-norm regularization of W is added to (2) to ensure that W is sparse in rows. Thus the optimization problem becomes

$$\min_{W} \sum_{i=1}^{l} \mathcal{L}(\mathbf{f}(\mathbf{x}_i), \mathbf{y}_i) + \alpha\|W\|_{2,1}. \tag{3}$$

This is a general formulation of multi-task feature selection [18]. In this formulation, the importance of an individual feature is evaluated by simultaneously considering multiple tasks. In this way, different tasks help each other to select features assumed to be shared across tasks.

In many practical applications, the number of labeled samples l is quite small, and thus the learned W is often unreliable. Considering the data samples may lie on a low-dimensional manifold embedding in a high dimensional space, we propose to utilize the large amount of unlabeled samples to help learning W under the theme of manifold regularization (MR) [1]. MR has been widely used for capturing the local geometry and conducting low-dimensional embedding. In MR, the data manifold is characterized by a adjacency graph, which explores the geometric structure of the compact support of the marginal distribution. The geometry is then incorporated as an additional regularizer to ensure that the solution is smooth with respect to the data distribution. In our method, the regularization term is given by

$$\begin{aligned} \frac{1}{2}\sum_{p=1}^{c}\sum_{i,j=1}^{n}(f_i^p - f_j^p)^2 A_{ij} &= \frac{1}{2}\sum_{i,j=1}^{n} A_{ij}(\mathbf{f}_i^T\mathbf{f}_i + \mathbf{f}_j^T\mathbf{f}_j - 2\mathbf{f}_i^T\mathbf{f}_j) \\ &= \mathrm{tr}(F^T(D-A)F) \\ &= \mathrm{tr}(F^T L F), \end{aligned} \tag{4}$$

where $F = [\mathbf{f}(\mathbf{x}_1), \mathbf{f}(\mathbf{x}_2), \ldots, \mathbf{f}(\mathbf{x}_n)]^T \in \mathbb{R}^{n\times c}$ is the predictions over all the data (labeled and unlabeled). Here, A is the adjacency graph constructed using all the data, and each element A_{ij} indicates the similarity between sample $\mathbf{x}_i$ and $\mathbf{x}_j$; D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} A_{ij}$ and $L = D - A$ is the graph Laplacian matrix. The regularization term guarantees that if two samples are similar in the feature space, then their predictions will be close. In this way, the data geometric structure existed in high dimensional space is preserved and the data distribution information is well explored.