Image and Vision Computing 54 (2016) 22–30

Contents lists available at ScienceDirect

Image and Vision Computing

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

Exploring canonical correlation analysis with subspace and structured sparsity for web image annotation☆

Liang Tao[a,*], Horace H.S. Ip[b], Aijun Zhang[a], Xin Shu[c]

[a]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
[b]Department of Computer Science, City University of Hong Kong, Hong Kong, China
[c]College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China

ARTICLE INFO

Article history:
Received 5 September 2014
Received in revised form 2 September 2015
Accepted 25 June 2016
Available online 19 July 2016

Keywords:
Canonical correlation
Image annotation
Subspace learning
Sparsity

ABSTRACT

Canonical correlation analysis (CCA) has been extensively exploited for modelling Internet multimedia. However, two major challenges are raised for the classical CCA. First, CCA frequently fails to remove noisy and irrelevant features. Second, CCA cannot effectively capture the correlation between semantic labels, which is especially beneficial for annotating web images. In this paper, we propose a new framework that integrates structural sparsity and low-rank shared subspace into the least-squares formulation of CCA. Under this framework, multiple label interactions can be uncovered by the shared common structure of the input space. Meanwhile, a few highly discriminative features can be decided via the structural sparse norm. Owing to the presence of non-smooth structured sparsity, a new efficient iterative algorithm is derived with guaranteed convergence. The empirical studies over several popular web image data collections consistently deliver the effectiveness of our new formulation in comparison with competing algorithms.

1. Introduction

With the proliferation of photo-taking devices and online picture sharing platforms, automatic image annotation gains growing interest in the multimedia understanding. However, such unlimited amount of Internet images pose a considerable challenge for their organization and retrieval. Even though there are several appealing annotation models, such as dependence maximization [1], multi-label linear discriminant analysis [2], Laplacian regularized annotation [3–5] and annotation by mining the label correlation graph structure [6], how to efficiently annotate these web images still remains a key research issue in the computer vision community.

One of the typical schemes for manipulating visual features with textual descriptions is built on canonical correlation analysis (CCA) which maximally measures the similarities between a pair of data features, i.e., the high dimensional input feature space and the dimensionality-reduced label space. Thanks to its desirable theoretical properties, CCA has been successfully leveraged in a large spectrum of multimedia analysis tasks, such as three-view cross modal retrieval [7], multi-modal hashing with correlation maximization [8] and feature-aware label space dimensionality reduction [9]. However, there are several inherent limitations of the standard CCA. Since dissimilar visual descriptors reflect distinct aspects of the visual characteristics, the input features commonly concatenate different visual descriptors, leading to high dimensional visual features in image annotation. From the perspective of feature selection, there are typically a small number of informative features for data analysis, but CCA cannot effectively remove the irrelevant/noisy features and sparsely select the informative ones from the high dimensional original features. In other words, the assumption of sparsity has been validated to promote better interpretation and generalization performance, but CCA cannot be expected to yield the attractive sparse representation of the projection matrix. Further, the standard CCA scales cubically with dimensionality due to the computation of generalized eigen-decomposition.

On the other hand, from the viewpoint of the semantic label dependence mining, CCA does not take into account the intrinsic interactions [1,10,11,12] among multiple pre-given labels that are quite helpful for the annotation problem. Although image annotation fundamentally concerns the classification problem, unlike the conventional binary classification, we cannot simply decompose the web image annotation into a series of independent binary classifications while ignoring the hidden correlation between semantic labels.

Inspired by the equivalence relationship [13] between the least-squares and the generalized eigen-value problem, we propose a

☆ This paper has been recommended for acceptance by Y. Aloimonos, PhD.

* Corresponding author. Tel. : +852 6849 8236.
E-mail addresses: liang.tao@my.cityu.edu.hk (L. Tao), cship@cityu.edu.hk (H. Ip), ajzhang@hkbu.edu.hk (A. Zhang), xinshu@outlook.com (X. Shu).

http://dx.doi.org/10.1016/j.imavis.2016.06.008
0262-8856/

new CCA model by simultaneously incorporating the shared common structure learning and row sparsity-inducing $\ell_{2,p}$-norm into a unified objective, dubbed shared subspace and structural sparse CCA (SSCCA$_{2,p}$). Specifically, we employ a row-wise structured sparsity regularizer which shrinks some rows of projection functions to zeros, to identify the essential discriminative features and eliminate the redundant and noisy dimensions for predictive functions; meanwhile, we exploit the shared common structure to encourage the interactions among different semantic labels so as to compensate for the CCA's lack of the semantic correlation captured in the embedded space. Under this scheme, not only can we elucidate the multiple label dependence unveiled by the shared structure, but also maximally characterize the similarity between the input feature space and the label space via CCA. Owing to the inclusion of the non-smooth row-sparsity term in this unified formulation, we derive an iterative alternating learning paradigm on the basis of the efficient randomization scheme to avoid the expensive exact eigen-decomposition in each iteration. By making use of the learned predictive classifiers, we have conducted extensive evaluations on different Web image corpora to showcase the competitive advantages of our model for efficient web image annotation.

2. Related work

To address the annotation issue, many recent methods have been proposed by capturing the label correlation. We begin with the multi-label dimensionality reduction via dependence maximization (MDDM). MDDM [1] utilizes the Hilbert–Schmidt Independence Criterion to maximize the dependence between the input space and the corresponding labels. MDDM, however, inefficiently adopts a kernel function for the label space to capture the correlation between multiple labels. Designing and hand-tuning appropriate kernel functions for different label spaces can be time-consuming and requires domain knowledge.

What is more, based on the incorporation of the normalized cosine similarity in the label matrix, multi-label linear discriminant analysis (MLDA) [2] extends the classical linear discriminant analysis, and thus facing the expensive generalized eigenvalue problem as well. Besides, a new graph structured sparsity model [6] for annotation leverages the element in the cosine label-wise similarity matrix to respectively scale the associated sparse regularizers, bringing about a more complicated objective function with very high computational cost. In concrete, the time complexity is $O(cd^3)$ at each iteration (see Table 1 for a list of important notations used in this paper).

Unlike the above methods which straightforwardly exploits the label correlation, some other previous works [10,11,12,14] demonstrate that the shared subspace is particularly helpful in mining the label dependence. Intuitively, the underlying subspace shared among multiple labels can be interpreted as the principal components of the prediction functions. We also borrow the idea of shared subspace to uncover the correlation among semantic labels. In our framework, however, the promising joint sparsity-inducing norm $\ell_{2,p}$ is tailored to conduct feature learning for the shared structure. Plus we can flexibly determine the value of p for the purpose of controlling the sparseness in the process of structural feature learning according to the unique data set.

Table 1
Some important notations.

Notations: descriptions
n: The number of training samples
d: The dimension of data points
c: The number of multiple semantic labels
r: The shared dimensionality
k: The rank of the training data points X, i.e. $k = rank(X)$
$X \in \mathbb{R}^{d\times n}$: The input space X
$L \in \mathbb{R}^{n\times c}$: The label space L
$M_{(c)}$: A matrix is built from the first c columns of the matrix M

Our approach also has connections to CCA-based learning. Motivated by the locally linear embedding, locality preserving CCA (LPCCA) [15] has been presented to incorporate the local neighborhood relationship into CCA. Nevertheless, LPCCA still requires an expensive eigen-decomposition step and it may even lose its discriminative ability if the original data space does not satisfy the underlying smooth manifold assumption (see the experimental outcome in Section 4.3). In addition to the locality preserving CCA, the literature provides two recent sophisticated models towards CCA, including multi-label output codes using CCA (OCCA) [16] and group-structured sparse CCA (GCCA) [17]. Within the error-correcting scheme, OCCA has two separate components: encoding and decoding. In the first encoding stage, OCCA directly uses the standard canonical output variates as the codewords that are employed for the subsequent classifer and regression training. In the second decoding stage, OCCA needs the polynomial time complexity to perform a mean field approximation for a predictive distribution on labels. On the other hand, GCCA [17] imposes overlapping group lasso penalty on CCA by virtue of the first-order optimization, which generally inherits the slow convergence rate and polynomial time in each iteration. On top of that, it could not be easy to determine these important and sensitive parameters such as the number of overlapping group and the weights of groups in GCCA. Because of their considerably high computational burdens and complicated parameters tuning, the aforementioned frameworks cannot efficiently handle large-scale web image data sets.

3. The SSCCA$_{2,p}$ framework

In this section we formulate the problem of the least-squares CCA under a new framework by simultaneously introducing the structural sparse norm and the common subspace extracted among different semantic labels. We first describe the shared structure that greatly assists the multi-label prediction, then present our new unified model tackled by an efficient alternating iterative optimization. Additionally, we use $X = \{x_1, \cdots, x_n\} \in \mathbb{R}^{d\times n}$ denoting the n training data points of dimension d and $L \in \{0,1\}^{n\times c}$ standing for the label space such that $L_i^j = 1$ if x_i is grouped into j-th label, and 0 otherwise, where c is the number of labels. Without any loss of generality, we consider both the input data space X and the label space L are normalized to have zero mean, i.e. $\sum_{i=1}^n X_{\cdot i} = 0$ and $\sum_{i=1}^n L_{i\cdot} = 0$.

3.1. Shared subspace with joint sparsity

Following the supervised learning framework, we aim to learn the projection functions $\left\{f_Q^j(x)\right\}_{j=1}^c$ and the low rank discriminative subspace Q from the input training data X by minimizing the below regularized empirical risk:

$$\min_{Q,f_Q^j} \sum_{j=1}^{c} \left(\sum_{i=1}^{n} \mathcal{F}(f_Q^j(x_i), y_i^j) + \lambda \mathcal{G}(f_Q^j) \right), \tag{1}$$

where y_i^j is a well-defined response variable, $\mathcal{F}(\cdot)$ is a prescribed loss function over the labeled data, the regularizer $\mathcal{G}(\cdot)$ measures the complexity of f_Q^j and the tradeoff regularization parameter λ controls the fitness of predictive functions. In order to capture the common subspace [10,11,12,14] shared among different semantic labels, we can define the predictive classifier

$$f_Q^j(x) = w_j^{\mathrm{T}} x = p_j^{\mathrm{T}} x + r_j^{\mathrm{T}} Q^{\mathrm{T}} x, \quad j = 1, \ldots, c, \tag{2}$$