Pattern Recognition 45 (2012) 2005–2018

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data

Xiaohong Chen [a,b], Songcan Chen [b,c,*], Hui Xue [d], Xudong Zhou [b]

[a] Department of Mathematics, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China
[b] Department of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China
[c] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China
[d] School of Computer Science and Engineering, Southeast University, Nanjing 210096, China

ARTICLE INFO

Article history:
Received 16 August 2011
Received in revised form
12 November 2011
Accepted 15 November 2011
Available online 25 November 2011

Keywords:
Multi-view data
Correlation analysis
Semi-supervised learning
Semi-paired learning
Dimensionality reduction

ABSTRACT

Canonical correlation analysis (CCA) is a popular and powerful dimensionality reduction method to analyze paired multi-view data. However, when facing semi-paired and semi-supervised multi-view data which widely exist in real-world problems, CCA usually performs poorly due to its requirement of data pairing between different views and un-supervision in nature. Recently, several extensions of CCA have been proposed, however, they just handle the semi-paired scenario by utilizing structure information in each view or just deal with semi-supervised scenario by incorporating the discriminant information. In this paper, we present a general dimensionality reduction framework for semi-paired and semi-supervised multi-view data which naturally generalizes existing related works by using different kinds of prior information. Based on the framework, we develop a novel dimensionality reduction method, termed as semi-paired and semi-supervised generalized correlation analysis (S^2GCA). S^2GCA exploits a small amount of paired data to perform CCA and at the same time, utilizes both the global structural information captured from the unlabeled data and the local discriminative information captured from the limited labeled data to compensate the limited pairedness. Consequently, S^2GCA can find the directions which make not only maximal correlation between the paired data but also maximal separability of the labeled data. Experimental results on artificial and four real-world datasets show its effectiveness compared to the existing related dimensionality reduction methods.

1. Introduction

In real world, we often meet with such a case that one object is represented by two or more types of features, e.g., gene can be represented by the genetic activity feature and text information feature [1], the same person has visual and audio features [2], each webpage can be represented by the text in the page and the hyperlinks jointly [3], CAD-catalogs are represented by some kind of 3D model like Bezier curves or polygon meshes and additional textual information like descriptions of technical [4]. This kind of data is usually called multimodal or multi-modality [1,2,5–8], multiple outlooks [9], multi-represented objection [4] or multi-view [3,10–14] data (for convenience, we will uniformly call them multi-view data hereafter). Analyzing such multi-view data to acquire useful information and knowledge has attracted more and more attentions recently. These works include dimensionality reduction (DR) [7,8,14–21], regression [22] and clustering [1,4,11]. In this paper, we focus on DR for multi-view data with the aim to avoid the curse of dimensionality [23] and overfitting brought by high dimensionality for good generalization [15], i.e., learning the appropriate low-dimensional representations for high dimensional data for subsequent task.

In recent years, a number of efficient algorithms [7,8,14–21] emerged to address this problem for discovering inherent structures and relations among different views. Among all the methods, canonical correlation analysis (CCA) [16–18] is the most widely used one. It works with two sets of related variables ($\mathbf{x}$, $\mathbf{y}$), and aims to find the directions that maximize the correlation between the two sets of projected representations in the low-dimensional space. In its implementation, CCA requires the data be rigorously paired or one-to-one correspondence among different views due to its correlation definition. However, such requirement is usually not satisfied in real life due to various reasons, e.g., (1) different sampling frequencies of sensors acquiring data or sensor faulty in an audio-video system, which result in

* Corresponding author at: Department of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China. Tel.: +86 25 84892452.
E-mail address: s.chen@nuaa.edu.cn (S. Chen).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.11.008

non-synchronicity between signals from different channels and even the missing of samples of certain views so that the multi-view data cannot keep one-to-one correspondence any more [7]. (2) Even having sufficient individual-view data, pairing them is still difficult, time consuming, even expensive since needing the efforts from experienced human annotators. Meanwhile, unpaired multi-view data are relatively easier to be collected. So we are often given only a few paired and a lot of unpaired multi-view data. We refer such data as semi-paired multi-view data. In literature, it is also named as weakly-paired multi-view data [6] or partially-paired multi-view data [19]. The common approaches to analyze such type of data include: (1) directly discarding unpaired data and performing correlation analysis just on the paired data, which usually results in overfitting on the given data and poor generalization for unseen samples especially when the paired data is scarce. (2) Creating synthetic samples in terms of certain criterion with the aim to generate paired multi-view data for correlation analysis [24]. These methods cannot achieve reasonable improvement due to not incorporating the prior information (such as clustering hypothesis and manifold hypothesis) of the data. Now, the key point to address this problem is how to utilize the meaningful prior information hidden in additional unpaired data. Most recently, some improved algorithms of CCA that can effectively deal with semi-paired multi-view data have emerged. Typically, Blaschko et al. [20] proposed semi-supervised Laplacian regularization of kernel canonical correlation (SemiLRKCCA) to find a set of highly correlated directions by exploiting the intrinsic manifold geometry structure of all data (paired and unpaired). Another paradigm is SemiCCA [15]. It essentially resembles the manifold regularization [25], i.e., using the global structure of the whole training data including both paired and unpaired samples to regularize CCA. Consequently, SemiCCA seamlessly bridges CCA and principal component analysis (PCA) [26,27], and inherits some characteristics of both PCA and CCA. It is necessary to mention that the actual meaning of "semi-" in SemiCCA and SemiLRKCCA is "semi-paired" rather than "semi-supervised" in popular semi-supervised learning literature [28,29]. Most recently, Gu et al. [19] proposed partially paired locality correlation analysis (PPLCA), which effectively deals with the semi-paired scenario of wireless sensor network localization by virtue of the combination of the neighborhood structure information in data. SemiCCA, SemiLRKCCA and PPLCA all cater well for semi-paired multi-view scenario and thus achieve better empirical results than CCA through preserving original paired information and deeply utilizing the structure information simultaneously.

As we have known, discriminative information is quite important for DR serving the classification task. However, SemiCCA, SemiLRKCCA and PPLCA are unsupervised DR methods, thus only concerning the between-view correlation embedded the structure information of each view is generally not enough for better classification accuracy. Concretely, SemiLRKCCA utilizes the graph Laplacians constructed through the unsupervised within-view k-nearest neighbors with regardless of the labeled or unlabeled data. SemiCCA employs unsupervised PCA as within-view regularization terms to do semi-paired learning. PPLCA replaces total mean with the neighborhood means into the formulation of CCA in each view such that PPLCA can incorporate the unpaired data information. Due to not exploiting the class information, the above three methods unavoidably result in the limitation of recognition performance. To overcome the limitation, Sun et al. [8] proposed the discriminative canonical correlation analysis (DCCA) for supervised multi-view data. DCCA aims to obtain DR with discrimination by maximizing the within-class correlation while minimizing the between-class correlation. Next, Sun et al. [7] further extended DCCA to the fully supervised and semi-paired scenario and developed the DCCA with Missing Samples (DCCAM). Borrowing the idea of DCCA, Peng et al. [21] proposed the local discrimination CCA(LDCCA) by incorporating the idea of local discriminant analysis [30] into CCA. Specifically, LDCCA takes local discriminant information of each view data into account for defining the local between-class covariance and local within-class covariance matrices and thus attempts to achieve effective between-class separation by maximizing local within-class correlations and minimizing local between-class correlations simultaneously. Essentially, the common key of above three methods is to construct the within-class and between-class correlation matrices. However, such a construction is only suit for the case that the class label-aligned discriminant information is given for each view data, hence their performance will degrade greatly when just few labeled data can be available.

Although DCCA, DCCAM and LDCCA can work reasonably well in fully supervised case, in many real-world applications such as image classification, web page classification and protein function prediction, labeled samples are harder to be collected than unlabeled samples since the labeling process is relatively expensive and time consuming. Thus, a semi-supervised(SSL) scenario occurred [28,29,31]. Recently, the multi-view DR in semi-supervised scenario has received increasing attention as a learning paradigm. For example, Foster et al. [22] performed CCA first for unlabeled data and then least squares regression for given labeled data in the CCA-generated lower dimensional subspace. Kursun and Alpaydin [32] proposed a Semi-supervised CCA(SCCA). In its implementation, a key ingredient is to rebuild two-view data and then perform correlation analysis, i.e., first for the one view, SCCA keeps the other view when class label is absent, otherwise replaces the samples by the corresponding class-centers, and then performs semi-supervised DR for this view data, the same process is repeated for the other view. Most recently, Hou et al. [14] developed a multiple view semi-supervised dimensionality reduction (MVSSDR) method with the discriminative information from given within-view pairwise must-link and cannot-link constraints (similar to SSDR [33]). Here a pair of "must-link" samples implies that they belong to the same classes of the same view and a pair of "cannot-link" samples implies that they belong to different classes of the same view. MVSSDR exploits the disparate structures and different statistical properties of different views to achieve better performance than SSDR which is only fit for all the concatenated representations of all the views. The above two methods [14,32] deal with a fully paired and semi-supervised multi-view case. Undoubtedly, such a strict pairing requirement among views naturally limits their applications in real world.

With the successive emergence of new application problems and the rapid development of data collection and processing techniques, multi-view data is more complex and diverse, i.e., between-view data may be paired or unpaired, and within-view data may be labeled or unlabeled simultaneously. According to whether the multi-view data under study is fully paired or not, the existing corresponding DR methods can be roughly categorized into paired ones (CCA, SCCA, MVSSDR, DCCA and LDCCA) and semi-paired ones (SemiCCA, SemiLRCCA, PPLCA and DCCAM). The former can further be divided into unsupervised, semi-supervised and supervised ones. The latter is subdivided into supervised and unsupervised ones. Table 1 summarizes the characteristics of the above related methods in terms of pairing information, discriminative information and structural information used.

From the "paired information" and "discriminative information" columns of Table 1, we observe that, there is no DR method to deal with semi-paired and semi-supervised multi-view data. Furthermore, we find that besides the paired information, both discriminative information and structural information are