Pattern Recognition 67 (2017) 313–327

Contents lists available at ScienceDirect

Pattern Recognition

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

Unified subspace learning for incomplete and unlabeled multi-view data

Qiyue Yin, Shu Wu, Liang Wang*

Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 11 August 2016
Revised 15 January 2017
Accepted 19 January 2017
Available online 17 February 2017

Keywords:
Multi-view learning
Subspace learning
Incomplete and unlabeled data
Multi-view clustering
Cross-modal retrieval

ABSTRACT

Multi-view data with each view corresponding to a type of feature set are common in real world. Usually, previous multi-view learning methods assume complete views. However, multi-view data are often incomplete, namely some samples have incomplete feature sets. Besides, most data are unlabeled due to a large cost of manual annotation, which makes learning of such data a challenging problem. In this paper, we propose a novel subspace learning framework for incomplete and unlabeled multi-view data. The model directly optimizes the class indicator matrix, which establishes a bridge for incomplete feature sets. Besides, feature selection is considered to deal with high dimensional and noisy features. Furthermore, the inter-view and intra-view data similarities are preserved to enhance the model. To these ends, an objective is developed along with an efficient optimization strategy. Finally, extensive experiments are conducted for multi-view clustering and cross-modal retrieval, achieving the state-of-the-art performance under various settings.

1. Introduction

Various kinds of real-world data appear in multiple modalities or come from multiple channels. For example, a web page can be described by both images and texts, and an image can be encoded by different visual features such as SIFT and GIST. Such data are called multi-view data with each view representing a type of feature set and these views can be homogeneous descriptors or heterogeneous modalities. Usually, multiple views provide complementary information for the semantically same data, which motivates the multi-view learning to obtain better performance than using a single view [1]. Besides, Multi-view data describing the same content lead to the research of exploring consistent information between different views, which results in cross-modal matching tasks [2].

Recently, plenty of methods have been developed for multi-view data to explore complementarity and consistency characteristics. It should be noted that most methods focus on complete multi-view data, which means all data samples in the datasets have complete feature sets. However, in real applications, it is often the case that some views suffer from missing information leading to incomplete multi-view data. For example, given a two view dataset with visual and textual features, some samples have only either visual or textual feature with only part of them sharing both representations. Under such scenario, traditional multi-view learning methods usually face notable performance degeneration [3,4]. Besides, real multi-view data are often unlabeled due to the expensive cost of manual annotation, which makes the learning of incomplete multi-view data a challenging problem.

Generally, to model incomplete and unlabeled multi-view data, we confront two basic challenges. The first one is how to handle incomplete multi-view data. Since some samples have incomplete feature representations, a naive strategy is to remove such examples and only use samples with complete feature sets. However, such methods are contradicting with some tasks such as clustering because we need to cluster all the data samples. More importantly, they cannot make full use of the whole data to learn models. Another strategy is to fill missing information. For example, matrix completion based methods [5] utilize low rank structure of the matrix to fill missing entities. However, those methods usually cannot perform feature selection to deal with high dimensional and noisy features. Thus by filling missing information is not a satisfactory strategy. Overall, a suitable model should use samples with complete feature representations and meanwhile utilize examples with incomplete feature sets to enhance the learning process.

The second challenge is how to explore complementarity and consistency for unlabeled multi-view data. Usually, for multi-view data describing semantically same content, different views

* Corresponding author.
E-mail addresses: qyyin@nlpr.ia.ac.cn (Q. Yin), shu.wu@nlpr.ia.ac.cn (S. Wu), wangliang@nlpr.ia.ac.cn (L. Wang).

http://dx.doi.org/10.1016/j.patcog.2017.01.035
0031-3203/

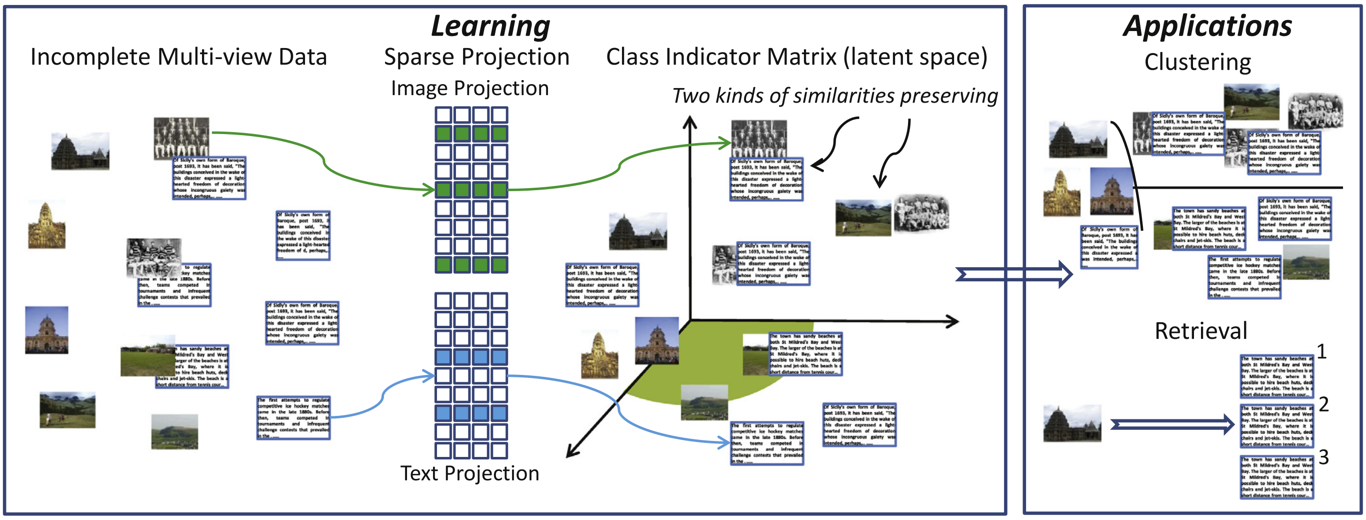

Fig. 1. The overview of our model with two views, i.e., text and image. For the incomplete multi-view dataset, we use projection matrix to project the original features to the class indicator matrix, which explicitly captures the clustering structure and serves as the latent space. Besides, group sparsity is imposed on the projection matrices for feature selection. Furthermore, the inter-view and intra-view data similarities are preserved to enhance the model. Finally, our model can be applied for clustering and retrieval tasks.

share common characteristics and have view-specific characteristics, which makes the modeling of those characteristics complex. Furthermore, given unlabeled data, we just have the corresponding relation between different views and this makes discover the structure of multi-view data harder. Most previous methods try to find a low dimensional subspace, where data samples under different views can be compared for exploring the above characteristics. For example, canonical correlation analysis (CCA) based approaches [6,7] aim to find linear projections of different views with maximal mutual correlation, and multi-view non-negative matrix factorization based methods learn unified latent representations among multiple sources of information [8,9]. However, those methods cannot thoroughly explore the data semantics in the learned subspace. To sum up, one good subspace should reflect such information and meanwhile make use of multiple views.

In this paper, we propose a novel subspace learning framework to alleviate the above problems, as shown in Fig. 1. We directly optimize the class indicator matrix as a shared subspace through linear projection matrices, which has two advantages: 1) establishing a bridge for different views based on their optimized labels whether the multi-view data are complete or incomplete, and 2) the class indicator matrix in turn guides the subspace learning in a supervised manner to make the learning process more accurate. Since data are often with high dimensional and noisy features, the projection matrices are enforced to be sparse to select relevant features when learning the latent space. Furthermore, the inter-view and intra-view data similarities are preserved to enhance the subspace learning. To these ends, an objective is developed with an efficient optimization strategy and convergence analysis. The experimental results show that our method outperforms the state-of-the-art methods.

Our contributions can be summarized as follows. 1) We propose a novel subspace learning based incomplete and unlabeled multi-view learning method, which jointly considers feature selection and inter-view and intra-view similarity preserving to enhance the subspace learning. 2) We develop an iterative optimization algorithm to efficiently solve the proposed objective, and provide theoretical analysis to guarantee its convergence. 3) We validate our proposed method with extensive experiments under two settings in terms of two tasks, i.e., multi-view clustering and cross-modal retrieval, achieving better performance than the state-of-the-art methods.

The rest of the paper is organized as follows. In Section 2, we briefly review multi-view learning, especially multi-view clustering and cross-modal retrieval. Section 3 describes our model, along with its optimization and convergence analysis. In Section 4, we report experimental results on multi-view clustering and cross-modal retrieval. Finally, we draw the conclusion in Section 5.

2. Related work

In this section, we briefly review general multi-view learning methods. Since we are focusing on two specific multi-view learning tasks, i.e., multi-view clustering and cross-modal retrieval, we also introduce recent progresses of them.

2.1. Multi-view learning

Multi-view learning deals with data represented by multiple distinct feature sets and aims at boosting learning performance or discovering correlation. It has a wide range of applications [56–58], such as dimensionality reduction, classification, retrieval and clustering. Generally, existing multi-view learning algorithms can be categorized into three schemes [1]. Co-training [10] is one of the earliest framework, which alternately maximizes the agreement of two feature sets. Soon after, plenty of variants are developed, such as generalized expectation-maximization (EM) and methods fusing co-training and other algorithms [11]. Multiple kernel learning solves multi-view learning by regarding different kernels as different views and then combining those kernels through linear or non-linear strategies. Such framework is widely studied and readers can refer to [12] for more details. The last framework is subspace learning, which aims to find a low dimensional space to measure the consistency and complementarity among multi-view data. Typical examples such as Canonical Correlation Analysis (CCA) and its various extensions [13–15] have obtained promising results in various tasks. In this paper, a novel subspace learning framework is developed for learning incomplete and unlabeled multi-view data.

2.2. Multi-view clustering

Multi-view clustering, as one of basic tasks of multi-view learning, provides a natural way to cluster multi-view datasets [16–18].