NeuroImage xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

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

Review

Bi-level multi-source learning for heterogeneous block-wise missing data

Shuo Xiang [a,b], Lei Yuan [a,b], Wei Fan [c], Yalin Wang [a], Paul M. Thompson [d], Jieping Ye [a,b,*], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA
[b] Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, Tempe, AZ, USA
[c] Huawei Noah's Ark Lab, Hong Kong
[d] Imaging Genetics Center, Laboratory of Neuro Imaging, Department of Neurology & Psychiatry, UCLA School of Medicine, Los Angeles, CA, USA

ARTICLE INFO

Article history:
Accepted 9 August 2013
Available online xxxx

Keywords:
Alzheimer's disease
Multi-modal fusion
Multi-source
Block-wise missing data
Optimization

ABSTRACT

Bio-imaging technologies allow scientists to collect large amounts of high-dimensional data from multiple heterogeneous sources for many biomedical applications. In the study of Alzheimer's Disease (AD), neuroimaging data, gene/protein expression data, etc., are often analyzed together to improve predictive power. Joint learning from multiple complementary data sources is advantageous, but feature-pruning and data source selection are critical to learn interpretable models from high-dimensional data. Often, the data collected has block-wise missing entries. In the Alzheimer's Disease Neuroimaging Initiative (ADNI), most subjects have MRI and genetic information, but only half have cerebrospinal fluid (CSF) measures, a different half has FDG-PET; only some have proteomic data. Here we propose how to effectively integrate information from multiple heterogeneous data sources when data is block-wise missing. We present a unified "bi-level" learning model for complete multi-source data, and extend it to incomplete data. Our major contributions are: (1) our proposed models unify feature-level and source-level analysis, including several existing feature learning approaches as special cases; (2) the model for incomplete data avoids imputing missing data and offers superior performance; it generalizes to other applications with block-wise missing data sources; (3) we present efficient optimization algorithms for modeling complete and incomplete data. We comprehensively evaluate the proposed models including all ADNI subjects with at least one of four data types at baseline: MRI, FDG-PET, CSF and proteomics. Our proposed models compare favorably with existing approaches.

Contents

* Corresponding author at: Department of Computer Science and Engineering, Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, 699 S. Mill Ave, Tempe, AZ 85287, USA.
E-mail address: jieping.ye@asu.edu (J. Ye).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but most of them did not participate in analysis or writing of this report. A complete listing of ADNI investigators may be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.08.015

Introduction

Alzheimer's Disease (AD), the most common form of dementia, is a highly prevalent neurodegenerative disease, in which memory and other cognitive functions decline gradually and progressively over time. AD accounts for 50–80% of dementia cases and the number of people affected by AD is expected to increase substantially over the coming decades (Brookmeyer et al., 2007). Currently there is no known cure for AD, but the detection and diagnosis of the onset and progression of AD in its earliest stages is invaluable and is the target of intensive investigation world-wide.

Recent advances in data collection technologies make it possible to collect a large amount of data to study and monitor the progression of AD. Often, these data come from multiple sources, and many studies involve multi-modality imaging. For example, different types of measurements based on magnetic resonance imaging (MRI) of the brain, positron emission tomography (PET), cerebrospinal fluid (CSF), blood tests, gene/protein expression data, and genetic data have been collected. These data are not redundant, and each of them provides complementary information for the diagnosis of AD (Calhoun et al., 2009; Fjell et al., 2010; Landau et al., 2010; Walhovd et al., 2010a). Extraction of the most useful information from such multi-source (i.e., multi-modality) data is critical in AD research. Data mining and machine learning methods have been increasingly used to analyze multi-source data (Calhoun et al., 2009; Crammer et al., 2008; Fan et al., 2008; Hinrichs et al., 2011; Troyanskaya et al., 2003; Vemuri et al., 2009; Walhovd et al., 2010b; Wang et al., 2012; Xu et al., 2007; Ye et al., 2008; Yuan et al., 2012; Zhang and Shen, 2012; Zhang et al., 2011). It is clear that both diagnostic and predictive power can be significantly improved if information from different sources is properly integrated and leveraged. Multi-source learning has thus attracted great attention in biomedical research (Calhoun et al., 2009; Huopaniemi et al., 2010; Ye et al., 2008). Multi-source learning is closely related to an area known as "multi-view" learning, but the two approaches differ in several important respects. More specifically, multi-view learning mainly focuses on semi-supervised learning and using unlabeled data to maximize the agreement between different views (Ando and Zhang, 2007; Culp et al., 2009). In this paper, we focus on multi-source learning in the supervised setting and we do not assume there are abundant unlabeled data available. In addition, we do not attempt to reduce the disagreement between multiple sources but try to extract complementary information from them, as is often the case in biomedical applications such as the study of AD.

In many applications including the study of AD, some of the available data also have a very high dimensionality, e.g., neuroimages or gene/protein expression data. However, this high-dimensional data often contains redundant information, as well as noisy or corrupted entries, and thus poses a potential challenge. To build a stable and comprehensive learning model with good generalization, it is common to apply *feature selection* – which identifies a small set of the most informative features – as a pre-processing step for classification or regression. One simple approach is to pool data from multiple sources together to create a single data matrix and apply traditional feature selection methods directly to the pooled data matrix. However, such an approach treats all sources as equally important, and ignores within-source and between-source relationships.

Another popular approach is to adopt multiple kernel learning (MKL) to perform data fusion (Lanckriet et al., 2004; Xu et al., 2007; Ye et al., 2008). This provides a principled method to perform source-level analysis, i.e., a particular source is considered relevant to the learning task only if its corresponding kernel is selected in the MKL approach. However, MKL only performs source-level analysis, ignoring feature-level analysis. Such an approach is suboptimal when the individual data sources are high-dimensional, and an interpretable model is desired. To fully take advantage of multi-source data, it is desirable to build a model that performs both individual feature-level and source-level analysis. In this paper, we will use the term "bi-level analysis", which was introduced in (Breheny and Huang, 2009), to refer to feature- and source-level analysis, performed simultaneously.

Besides the multi-modality aspects and the high dimensionality of the data, a further problem is very commonly encountered: the existence of (block-wise) missing data is another major challenge encountered in AD and other biomedical applications. Fig. 1 provides an illustration of how block-wise missing data arises in AD research. In this example, we have 245 participants in total and 3 types of measurements (PET, MRI and CSF) represented in different colors. The blank region means that data from the corresponding source is missing. In this example, participants 1–139 have available data for PET and MRI but lack CSF information, while participants 149–245 have only MRI data. The block-wise missing data situation tends to emerge in several scenarios: low-quality data sources of certain samples may be discarded; some data-collecting mechanisms (like PET) may be too costly to apply to every participant; participants may not be willing to allow certain measurements, for various reasons (e.g., lack of consent, contraindications, participant attrition, non-compliance with a long scan). Note that the missing data often emerges in a block-wise fashion, i.e., for a patient, a certain data source is either present or missing completely.

Considerable efforts have been made to deal with missing data, both in the data mining and neuroimaging communities. Some well-known missing value estimation techniques like EM (Duda et al., 1997), iterative singular value decomposition (SVD) and matrix completion (Mazumder et al., 2010) have been extended to biomedical applications by performing *imputation* on the missing part of the data. Although these approaches are effective in handling random missing entries, they often deliver sub-optimal performance in AD research (Yuan et al., 2012) for the following reasons: (1) these imputation approaches fail to capture the pattern of the missing data, i.e., the missing elements are not randomly scattered across the data matrix but emerge block-wise. However, such prior knowledge is completely discarded in imputation methods; (2) due to the high dimensionality of the data, these methods often have to estimate a significant amount of missing values, which can lead to unstable performance.

To overcome the aforementioned drawbacks of standard imputation methods, we previously proposed an incomplete Multi-Source Feature learning method (iMSF) which avoids direct imputation (Yuan et al., 2012). The iMSF method first partitions the patients into disjoint groups, so that patients from the same group possess identical data source combinations. Feature learning is then carried out independently in each group and finally the results from all groups are appropriately combined to obtain a consistent feature learning result. Such a mechanism enables iMSF to perform feature selection without estimating the

Please cite this article as: Xiang, S., et al., Bi-level multi-source learning for heterogeneous block-wise missing data, NeuroImage (2013), http://dx.doi.org/10.1016/j.neuroimage.2013.08.015