NeuroImage 91 (2014) 386–400

Contents lists available at ScienceDirect

NeuroImage

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

Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion

Kim-Han Thung [a], Chong-Yaw Wee [a], Pew-Thian Yap [a], Dinggang Shen [a,b,*], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Biomedical Research Imaging Center (BRIC) and Department of Radiology, University of North Carolina at Chapel Hill, USA
[b] Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

ARTICLE INFO

Article history:
Accepted 18 January 2014
Available online 27 January 2014

Keywords:
Matrix completion
Classification
Multi-task learning
Data imputation

ABSTRACT

In this work, we are interested in predicting the diagnostic statuses of potentially neurodegenerated patients using feature values derived from multi-modality neuroimaging data and biological data, which might be incomplete. Collecting the feature values into a matrix, with each row containing a feature vector of a sample, we propose a framework to predict the corresponding associated multiple target outputs (e.g., diagnosis label and clinical scores) from this feature matrix by performing matrix shrinkage following matrix completion. Specifically, we first *combine the feature and target output matrices into a large matrix and then partition this large incomplete matrix into smaller submatrices*, each consisting of samples with complete feature values (corresponding to a certain combination of modalities) and target outputs. Treating each target output as the outcome of a prediction task, we apply a *2-step multi-task learning algorithm to select the most discriminative features and samples* in each submatrix. Features and samples that are not selected in any of the submatrices are discarded, resulting in a *shrunk version of the original large matrix*. The missing feature values and unknown target outputs of the shrunk matrix is then completed simultaneously. Experimental results using the ADNI dataset indicate that our proposed framework achieves higher classification accuracy at a greater speed when compared with conventional imputation-based classification methods and also yields competitive performance when compared with the state-of-the-art methods.

Introduction

Alzheimer's Disease (AD) is the most prevalent form of dementia. It is ultimately fatal and is ranked as the sixth leading cause of death in the United States in year 2012 (Alzheimer's Association, 2013). Neurodegeneration associated with AD is progressive and the symptoms usually begin with gradual memory decline followed by a gradual loss of cognitive and motor abilities that will cause difficulties in the daily lives of the patients. Eventually, the patients will lose the ability to take care of themselves and will need to rely on the intensive care provided by others. This has posed significant medical and socioeconomic challenges to the community (Alzheimer's Association, 2013).

Owing to the criticality of this issue, it is vital to diagnose AD accurately, especially at its prodormal stage, i.e., amnestic mild-cognitive impairment (MCI), so that an early treatment can be provided to possibly stop or slow down the progression of the disease. MCI, which is defined as a condition where the patient has noticeable cognitive decline, but without difficulty in carrying out daily activities, has high probability to develop into AD. With the help of emerging neuroimaging technology, the progress and severity of the neurodegeneration associated with AD or MCI can now be diagnosed and monitored in different ways (modalities). Magnetic resonance imaging (MRI) scans, for instance, provide 3D structural information about the brain, where features such as region-of-interest (ROI)-based volumetric measure and the cortical thickness can be extracted from the MRI to quantify brain atrophy that is usually associated with the diseases (Cuingnet et al., 2011; Desikan et al., 2009; Du et al., 2007; Fan et al., 2007b; Gerardin et al., 2009; Klöppel et al., 2008; Oliveira et al., 2010). Flourodeoxyglucose positron emission tomography (FDG-PET), on the other hand, can be used to detect abnormality in term of glucose metabolic rate at brain regions preferentially affected by AD (Chételat et al., 2003, 2005; Foster et al., 2007; Herholz et al., 2002; Higdon et al., 2004). Besides neuroimaging techniques, another line of research uses biological and genetic data to develop potential biomarkers for AD diagnosis. The important measurements in biological and genetic data that are closely related to cognitive decline in AD patients include the increase of cerebrospinal fluid (CSF) total-tau (t-tau) and CSF tau hyperphosphorylated

* Corresponding author.
E-mail addresses: khthung@email.unc.edu (K.-H. Thung), dgshen@med.unc.edu (D. Shen).

[1] Data used in the 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 did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

http://dx.doi.org/10.1016/j.neuroimage.2014.01.033
1053-8119/

at threonine 181 (p-tau), the decrease of CSF amyloid β ($\boldsymbol{A}\beta$), and the presence of gene apolipoprotein E (APOE) $\epsilon 4$ allele (Fagan et al., 2007; Fjell et al., 2010; Morris et al., 2009).

Although it is common to use information from only one modality such as structural MRI for diagnosis of AD/MCI, complementary information from multiple modalities (Fjell et al., 2010; Walhovd et al., 2010; Landau et al., 2010; Zhang et al., 2011; Liu et al., 2014; Verma et al., 2005; Fan et al., 2007a; Wee et al., 2011, 2012; Li et al., 2012; Zhou et al., 2011) can be combined for more accurate diagnosis. This is supported by the results reported in recent studies (De Leon et al., 2006; Fan et al., 2008; Ye et al., 2008; Hinrichs et al., 2009, 2011; Davatzikos et al., 2011; Zhang and Shen, 2012; Zhang et al., 2011; Liu et al., 2012; Zhang et al., 2012). To support AD research using multi-modality data, Alzheimer's Disease Neuroimaging Initiative (ADNI) has been actively collecting data from multiple modalities (e.g., MRI, FDG-PET and CSF data) from AD, MCI and normal control (NC) subjects yearly or half-yearly. Unfortunately, not all the samples in ADNI dataset are completed with the data from all different modalities. For example, while all the samples in the ADNI baseline dataset contain MRI data, only about half of the samples contain FDG-PET data (which is referred to as PET throughout the manuscript) and another different half of the samples contain CSF data. The "missing" data in the ADNI dataset is due to several reasons, such as, high measurement cost (i.e., PET scans), poor data quality and unwillingness of the patients to receive invasive tests (i.e., collection of CSF samples through lumbar puncture).

There are basically two approaches to deal with missing data in a dataset, i.e., we can either 1) discard the samples with missing data, or 2) impute the missing data. Most existing approaches discard samples with at least one missing modality and perform disease identification based on the remainder of the dataset. However, this approach discards a lot of information that is potentially useful. In fact, in following this approach for multi-modality analysis using MRI, PET and CSF data, about 2/3 of the total samples at ADNI baseline dataset will have to be removed.

The data imputation approach, on the other hand, is more preferable as it provides the possibility to use as many samples as possible in analysis. In fact, incomplete dataset is ubiquitous in many applications and thus various imputation methods have been developed to estimate the missing values based on the available data (Schneider, 2001; Troyanskaya et al., 2001; Zhu et al., 2011). However, these methods work well only when a small portion of the data is missing, but become less effective when a large portion of the data is missing (e.g., PET data in ADNI). Recently, low rank matrix completion (Candès and Recht, 2009) has been proposed to impute missing values in a large matrix through trace norm minimization. This algorithm can effectively recover a large portion of the missing data if the ground truth matrix is low rank and if the missing data are distributed randomly and uniformly (Candès and Recht, 2009). Unfortunately, the latter assumption does not hold in our case since, for each subject, the data from one or more of the modalities might be entirely missing, i.e., the data is missing in blocks.

In this paper, we attempt to identify AD and MCI from the NCs by using incomplete multi-modality dataset from the ADNI database. Denoting the incomplete dataset as a matrix with each row representing a feature vector derived from multi-modality data of a sample, conventional approach for solving this problem is to impute the missing data and build a classifier based on the completed matrix. However, it is too time consuming (as matrix size is large) (Jollois and Nadif, 2007; Xu and Jordan, 1996) and inaccurate (as there are too many missing values) to apply the current imputation methods directly. In addition, the errors introduced during the imputation process may affect the performance of the classifier. In this paper, we largely avert the problems of the conventional approach by proposing a framework (Thung et al., 2013) that 1) shrinks the large incomplete matrix through feature and sample selections, and 2) predicts the output labels directly through matrix completion on the shrunk matrix (i.e., without building another classifier on the completed matrix).

Specifically, we first partition the incomplete dataset into two portions – training set and testing set. Each set is represented by an incomplete feature matrix (each row contains feature vector of a sample), and a corresponding target output matrix (i.e., diagnostic status and clinical scores). Our first goal is to remove redundant/noisy features and samples from the feature matrix so that the imputation problem can be simplified. However, due to the missing values in the feature matrix, feature and sample selections cannot be performed directly. We thus partition the feature matrix, together with the target output matrix, into submatrices with only complete data (Ghannad-Rezaie et al., 2010), so that a 2-step multi-task learning algorithm (Obozinski et al., 2006; Zhang and Shen, 2012) can be applied to these submatrices to obtain a set of discriminative features and samples. The selected features and samples then form a shrunk, but still incomplete, matrix which is more "friendly" to imputation algorithms, as redundant/noisy features and samples have been removed and there are now a smaller number of missing values that need to be imputed. We propose to impute the missing feature data and target outputs simultaneously using a matrix completion approach. Two matrix completion algorithms are explored: low rank matrix completion and expectation maximization (EM). Experimental results demonstrate that our framework yields faster imputation and more accurate prediction of diagnostic labels than the conventional imputation-based classification approach.

In brief, we propose a framework for a solution for this problem – classification using incomplete multi-modality data with large block of missing data. The contributions of our framework are summarized below:

- Feature selection using incomplete matrix (i.e., matrix with missing values) through data grouping and multi-task learning.
- Sample selection using incomplete matrix through data grouping and multi-task learning.
- Improve imputation effectiveness by focusing only on the imputation of important data.
- Improve classification performance by label imputation.

Data

ADNI background

Data used in the preparation of this article were obtained from the ADNI database (`adni.loni.ucla.edu`). ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date, these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see `www.adni-info.org`.