Neurocomputing 195 (2016) 50–55

Contents lists available at ScienceDirect

Neurocomputing

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

Modeling and predicting AD progression by regression analysis of sequential clinical data

Qing Xie [a], Su Wang [a], Jia Zhu [b], Xiangliang Zhang [a,*], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] Division of CEMSE, King Abdullah University of Science and Technology, Saudi Arabia
[b] School of Computer Science, South China Normal University, China

ARTICLE INFO

Article history:
Received 6 April 2015
Received in revised form
19 June 2015
Accepted 16 July 2015
Available online 23 February 2016

Keywords:
Alzheimer's disease
Medical image
Regression
Sequential data analysis

ABSTRACT

Alzheimer's Disease (AD) is currently attracting much attention in elders' care. As the increasing availability of massive clinical diagnosis data, especially the medical images of brain scan, it is highly significant to precisely identify and predict the potential AD's progression based on the knowledge in the diagnosis data. In this paper, we follow a novel sequential learning framework to model the disease progression for AD patients' care. Different from the conventional approaches using only initial or static diagnosis data to model the disease progression for different durations, we design a score-involved approach and make use of the sequential diagnosis information in different disease stages to jointly simulate the disease progression. The actual clinical scores are utilized in progress to make the prediction more pertinent and reliable. We examined our approach by extensive experiments on the clinical data provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI). The results indicate that the proposed approach is more effective to simulate and predict the disease progression compared with the existing methods.

1. Introduction

Alzheimer's Disease (AD) is a kind of presenile dementia, which is the common neurodegenerative disease appearing in the elders over 65 years old, with the symptoms of memory loss and disorder of central nervous system, and potentially resulting in death [1]. It has been reported that there are over 26 million AD patients all over the world by 2011, and this number will go beyond 114 million by 2050 [2,3]. Therefore, the timely AD diagnosis and treatment are of high significance and have attracted much concern recently, and researchers have taken efforts to simulate and predict the disease progression to benefit the elders' care.

With the increasing availability of medical diagnosis data [4] and the development of image processing [5], machine learning methods have been engaging the AD pattern analysis and progression prediction based on the massive diagnosis data, especially the medical images of brain scan, including Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET). These neuroimaging data are popularly used to understand the AD progression and identify the diagnosis of AD and its early stage, Mild Cognitive Impairment (MCI). Our work will focus on the challenges and innovations in neuroimaging analysis.

There are several research directions for analyzing the AD progression based on the medical image data. The first one starts from the image processing direction and focuses on the volume of different brain regions [6]. Guo et al. [7] discovered the relationship between the AD progression and the decrease of gray matter volume in the hippocampus, parahippocampal gyrus and insula and superior temporal gyrus. However, such methods only limit the applications within individual levels and it is hard to discover more general rules to identify the progression. Besides this, many works propose to develop classification models to distinguish different disease status [8–10], and determine the clinically defined categories of the subjects [11,12], such as AD, MCI and healthy Normal Control (NC). Survival model has also been applied to simulate the AD progression from the statistical point of view [13,14]. Recently, as the clinical scores are commonly accepted to indicate the disease status, regression model has been more popularly investigated to predict the AD patients' status in terms

* Corresponding author.
E-mail addresses: qing.xie@kaust.edu.sa (Q. Xie), su.wang@kaust.edu.sa (S. Wang), jzhu@m.scnu.edu.cn (J. Zhu), xiangliang.zhang@kaust.edu.sa (X. Zhang).

[1] Data used in preparation of this paper were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.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 analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

http://dx.doi.org/10.1016/j.neucom.2015.07.145
0925-2312/

of clinical scores, such as Mini-Mental State Examination (MMSE) or Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog), based on the original diagnosis data [15–17]. In the regression model research, since the high-dimensional neuroimage data are involved to generate the features, various methods have been proposed to effectively improve the performance, such as dimension reduction technique [18,19] and feature selection technique [20–23]. Specifically, Zhang et al. [24] proposed to address both disease diagnosis and clinical score prediction simultaneously, and combine these two tasks in a unified framework based on the correlated feature selection, which is also called Joint Regression and Classification (JRC) problem [25]. From the prediction effectiveness point of view, most methods of the regression model aim to predict the target score at a specific time point, such as one year [18], while more prediction scores at different time points are desired for a better prediction performance.

To address this problem, multi-task learning techniques [26] have been introduced into the regression model to simulate the disease progression and predict the clinical scores at different time points [27,28]. Multi-task learning aims to improve the performance of regression model building by utilizing the intrinsic commonality among different target tasks. The shared representation in parallel learning can help individual tasked be learned better. It has been demonstrated that multi-task learning is especially effective when the number of subjects is small and the number of input features is large, which is the case of AD simulation. The essential issue of multi-task learning is to discover how the tasks are related and identify the learning model. To achieve this, Zhou et al. [27,28] model the problem as longitudinal disease progression and the predictions of a patient's disease status at different time points are treated as regression tasks. These prediction tasks for different time points are performed simultaneously and the temporal smoothness across prediction models can be captured. Specifically, their work develop formulations that allow the simultaneous selection of a common set of biomarkers for multiple time points and specific sets of biomarkers for different time points, so as to capture the temporal patterns of the biomarkers in disease progression.

The approaches based on multi-task learning model tactfully fuse the regression tasks for different time durations into the uniformed framework. Although different kinds of loss functions have been employed to refine the prediction model, they still have the following two drawbacks. Firstly, most works only learn the regression model based on the subject features generated at the baseline time point, and all the prediction analysis for the following time points is derived from the baseline information. Therefore, if the subjects for model learning only cover limited status of AD progression, e.g., the early stage, the regression model will be inaccurate to describe the progression. Secondly, these approaches fail to use the evolving feature information in progress, which can be also utilized as feedback to enhance the regression model. For a long term task, which is the case of the engaged problem, the actual evolution of the subjects in progress can increase the uncertainty for prediction and should not be ignored.

Therefore, we propose a sequential data analysis mechanism to perform the regression model for simulating the AD progression, and the feedback concept is involved in the model to improve the prediction performance. In our work, we will effectively make use of the intermediate information during the AD patients' therapy, which includes the consecutive brain scan images and corresponding clinical scores. Based on these information, our work will jointly analyze the feature data in different time points, and the existing clinical scores will be embedded into the model to facilitate the disease simulation and the clinical score prediction. The relationship is established by building a fused sparse Lasso formulation [29], which incorporates the temporal smoothness.

Generally speaking, our framework focuses more on the solution suitable for the specific application in medical care. It explores the information and resources available in practical conditions, and suggests the regression model accordingly. The idea of sequential analysis is an effective simulating approach for the continuous medical monitoring, as well as other similar scenarios.

To evaluate the performance of our approach, we conducted extensive experimental studies on the clinical image data. Data used in the preparation of this paper were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). The results evidently demonstrate the effectiveness and accuracy of our proposed solutions.

2. Regression by sequential diagnosis data analysis

In this section, we will interpret our solutions in detail. We will introduce the basic concepts and notations first, and then explain the regression model building by sequential data analysis.

2.1. Preliminaries

In the AD caring, the target patients will receive regular MRI or PET scan in fixed time interval, and their cognitive scores will be measured accordingly. Based on the acquired medical image data, the regression object aims to predict the cognitive score at specific time points.

Assume that a patient's brain-scan image can be processed into d-dimensional feature data, and for this patient' brain-scan records at all time points, we can collect all the feature data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Here, for each sample $\mathbf{x}_t \in \mathbb{R}^d$, there is a corresponding clinical score y_t (MMSE or ADAS-Cog) measured at the same time. In our test corpora, we have the record collection for all patients under monitoring. The regression model simulates the relationship between the collected feature data and the corresponding target clinical measures, so as to predict the patient's potential clinical score at the specific data point in future.

2.2. Regression model building

In this part, we will introduce three different approaches to build the feasible regression model based on sequential analysis. We will focus on the analysis of one patient to interpret the model building.

2.2.1. Baseline sequential prediction

The first model predicts the target clinical scores at different time points based on the baseline observation. When the patient takes the first brain scan, we can have his initial feature observation $\mathbf{x}_1$, and we predict his future clinical scores y_t $(t > 1)$ based on $\mathbf{x}_1$. Importantly, when the actual score y_t is available, we embed it into the baseline feature data to facilitate the prediction of y_{t+1}.

More specifically, assume that $\tilde{y}_t$ is the prediction (or estimation) of y_t, and the regression model can be defined as given the baseline observation $\mathbf{x}_1$, and the actual clinical score y_t $(t \geq 1)$, predict the estimation for the clinical score at next time point $\tilde{y}_{t+1}$. Such model is called Baseline Sequential Prediction (BSP).

The model of BSP is demonstrated in Fig. 1. From the figure, we can see that at each time point, the clinical score is predicted using