Neurocomputing 195 (2016) 104–111

Contents lists available at ScienceDirect

Neurocomputing

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

$L_{2,p}$−norm and sample constraint based feature selection and classification for AD diagnosis

Mingxing Zhang [a], Yang Yang [a,*], Hanwang Zhang [b], Fumin Shen [a], Dongxiang Zhang [b]

[a] University of Electronic Science and Technology of China, No.2006, Xiyuan Ave, West Hi-Tech Zone, 611731 Chengdu, China
[b] National University of Singapore, 21 Lower Kent Ridge Road, 119077, Singapore

ARTICLE INFO

Article history:
Received 21 March 2015
Received in revised form
26 June 2015
Accepted 23 August 2015
Available online 1 April 2016

Keywords:
Alzheimer's Disease (AD)
Mild Cognitive Impairment (MCI)
$L_{2,1}$−norm
Sparse learning
Multi-modality learning
Multi-task learning

ABSTRACT

Recent studies have witnessed the effectiveness of $L_{2,1}$-norm based methods on AD/MCI diagnosis. Nonetheless, most of them suffer from the following three main problems: (1) $L_{2,1}$−norm based loss function does not take into account different distances between target labels and prediction values; (2) $L_{2,1}$−norm based feature selection does not possess sufficient flexibility to adapt to different types of data sources and select more informative features; (3) intrinsic correlation between the processes of feature selection and classification (or regression) are inevitably ignored. In this paper, we propose a novel method which incorporates additional flexibility and adaptability by employing the more generalized $L_{2,p}$−norm based prediction loss function and $L_{2,q}$−norm based feature selection, as well as utilizes a joint model to perform feature selection and classification simultaneously. Besides, we introduce a regularizer to preserve local structure information between samples in the original feature space and prediction values in the projected space. In order to validate the effectiveness of the proposed method, we conducted extensive experiments on the ADNI dataset, and showed that the proposed method enhanced the performance of disease status classification, compared to the state-of-the-art methods.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Alzheimer's disease (AD) is the most common dementia in elderly people, which results in serious intellectual problems of memory, thinking and behavior. According to the report of Brookmeyer et al. [1], there would be over 30 million people around the world suffering from this disease by 2050. Its prodromal stage, which is called Mild Cognitive Impairment (MCI), can also lead to cognitive changes and high risk of development of AD over times [2]. It is very important to diagnose the AD/MCI and many researches about automatic computer-aid diagnosis of these diseases have been conducted in recent decades.

One of the main problems in the field of automatic medical diagnosis is that the dimension of medical data is normally far larger than the sample size. For example, the size of samples in many researches such as [3–5] was very small (only 103 samples with 51 for AD and 52 for NC), while the dimension of data features such as MRI and PET reached hundreds to thousands. However, AD is only related to a few areas of brain according to the research in [5]. These high-dimensional features usually contain many uninformative features. The high dimension of data also could result in the over-fitting problem [6] and the small size of samples makes it more serious.

To address these issues, feature selection based methods have been commonly used in literatures. As the successful applications of sparse method such as [7–11], $L_{2,1}$−norm based methods have been widely used in feature selection process for AD diagnosis. Wang et al. [12] proposed a multi-task learning method that performed the label classification and cognitive measure scores regression simultaneously. Different from traditional methods that selected features only associated with cognitive measure scores or disease status, this method selected the features related to both of them. Zhang et al. [3] proposed a multi-modal multi-task learning method that firstly selected the subset of features using Multi-Task method from each modality, then used the multi-modal support vector for the classification of AD and MCI. However, these methods do not consider the relationship between target vectors of samples. To overcome this disadvantage, Liu et al. [4] proposed a graph-matching feature selection method that preserves the relationship between the predicted vectors and the target vectors and takes high-order graph-matching. Furthermore, Zhu et al. [13] proposed a new loss function based on matrix-similarity that not only considers the natural relationship of clinical scores and label, but also the spatial relationship of samples to take a better feature selection and classification. This method was also developed in [14]. In order to use multi-modal data more effectively, Shi et al. [15] proposed a method that fuses the features from different modalities by using the

* Corresponding author.
E-mail address: dlyyang@gmail.com (Y. Yang).

http://dx.doi.org/10.1016/j.neucom.2015.08.111
0925-2312/© 2016 Elsevier B.V. All rights reserved.

pairwise coupled-diversity correlation. However, those previous methods based on $L_{2,1}-$norm regularizer suffer from some of three main disadvantages: (1) $L_{2,1}-$norm based loss function does not take into account different distances between target labels and prediction values; (2) $L_{2,1}-$norm based feature selection does not possess sufficient flexibility to adapt to different types of data sources and select more informative features; (3) intrinsic correlation between the processes of feature selection and classification (or regression) is inevitably ignored in existing methods.

In this paper, we propose a novel loss function that combines the $L_{2,p}-$norm of prediction loss function and $L_{2,q}-$norm of feature selection, and utilizes a joint model to conduct the feature selection and classification simultaneously. We also introduce a new item that keeps local structure information between samples in the feature space and prediction values in the projected space. The $L_{2,p}-$norm of prediction loss function attempts to adjust distances between predict values and target labels, and controls distances at the point of convergence of loss function. The larger p is, the less widely the distances vary. The $L_{2,q}-$norm of feature selection tries to control the sparsity of feature selection. The larger q is, the less sparse the feature selection is. We can flexibly select appropriate p, q according to the data and thus achieve a better classification. The new spatial information item preserves relationships between samples, *i.e.* if two samples are close in original feature space, they are still close neighbors in the projected space. Experiments on the ADNI dataset have showed that our proposal indeed helps us to enhance the performance of disease status classification, comparing the state-of-the-art methods.

2. Materials and preprocessing

In this paper, we use the ADNI (Alzheimer's Disease Neuroimaging Initiative) dataset[1] to evaluate our method. 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 Alzheimer's disease (AD). These data contains 202 samples including 51 samples for AD, 52 samples for NC, and 99 samples for MCI. We downloaded the MRI, PET and CSF data from the public ADNI website. Then we extract 93 features from MRI and 93 features from PET as well as 3 features from CSF following the widely used procedures such as in [3,16]. The detailed statistic information of these samples is showed in Table 1. Note that the numbers in this table represent the subjects' number or values' range corresponding to each category. MCI-C represents MCI Converters and MCI-NC represents MCI Non-Converters.

Table 1
Statistic information of samples in our dataset.

Items	AD (51)	NC (52)	MCI-C (43)	MCI-NC (56)
Female/male	18/33	18/34	15/28	17/39
Age	75.2 ± 7.4	75.3 ± 5.2	75.8 ± 6.8	74.8 ± 7.1
Education	14.7 ± 3.6	15.8 ± 3.2	16.1 ± 2.6	15.8 ± 3.2
MMSE	23.8 ± 2.0	29.0 ± 1.2	26.6 ± 1.7	27.5 ± 1.5
ADAS-Cog	18.3 ± 6.0	7.4 ± 3.2	12.9 ± 3.9	10.2 ± 4.3

3. Our method

In this section, we describe our feature selection and classification framework for AD/MCI diagnosis. Given the MRI, PET, and CSF features, we construct the feature matrix with each column concatenates these multi-modal features, and the target matrix or vector of ground truth with each column representing a sample that concatenates a class label and two clinical scores (ADAS-Cog and MMSE) or contains only a class label.

3.1. Preliminaries

Let $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_n]\in\mathbf{R}^{\mathbf{d}\times\mathbf{n}}$ and $\mathbf{Y}=[\mathbf{y}_1,\ldots,\mathbf{y}_n]^T\in\mathbf{R}^{\mathbf{c}\times\mathbf{n}}$, where n, d and c denote the number of samples, dimension of features, and number of target values, respectively. In our work, the target values correspond to a class label and two clinical scores, thus c can be equal to 1, 2 or 3. Commonly, we can predict the target variables by preforming a linear transformation for features, which is formulated as the following equation:

$$f(\mathbf{X})=\mathbf{X}^T\mathbf{W}=\hat{\mathbf{Y}} \tag{1}$$

where $\mathbf{W}\in\mathbf{R}^{d\times c}$ is a regression matrix and $\hat{\mathbf{Y}}$ is the predict target matrix. Each column in $\hat{\mathbf{Y}}$ corresponds to one of target variable and each row in $\hat{\mathbf{Y}}$ corresponds to one of samples. Note that, in this paper, $\mathbf{X}$ has been appended one additional dimension with value of 1 for every sample to include the bias item. Like the proposal of other literatures, if we restrict to select the same features to predict all target variables, we can formulate the feature selection method as follows:

$$\min_{\mathbf{W}} f(\mathbf{W})+\lambda\|\mathbf{W}\|_{2,1} \tag{2}$$

where $f(\mathbf{W})$ is the loss function between prediction values and target values depending on $\mathbf{W}$ and $\|\mathbf{W}\|_{2,1}=\sum_{i=1}^{d}\|\mathbf{W}^i\|_2$, $\mathbf{W}^i$ is the ith row of $\mathbf{W}$, λ is a weight parameter. The $L_{2,1}-$norm regularizer $\|\mathbf{W}\|_{2,1}$ let the model simultaneously select or not select a feature for predicting all target variables. Specifically, the L_2-norm regularizer in each row of $\mathbf{W}$ enforces all tasks to select the same features, and the L_1-norm regularizer imposes the constraint of sparseness in the feature selecting stage to select the most important and discriminative features. In our classification problem, our $L_{2,q}-$norm is like this $L_{2,1}-$norm but with additional flexibility and adaptability.

The loss function $f(\mathbf{W})$ in Eq. (2) is commonly defined as the distance between target values $\mathbf{Y}$ and predicted values of all samples, which is presented as follows:

$$\begin{aligned} f(\mathbf{W}) &= \|\mathbf{Y}-\mathbf{X}^T\mathbf{W}\|_F^2 \\ &= \|\mathbf{Y}-\hat{\mathbf{Y}}\|_F^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{c}(\mathbf{y}_{ij}-\mathbf{y}_{ij})^2 \end{aligned} \tag{3}$$

This distance based similarity metric has been proved effective and efficient in many literatures such as [3,13,17]. The lower this function value is, the more accurate the prediction is.

3.2. The proposed method

As described in the introduction, most $L_{2,1}-$norm based methods have three main disadvantages. One is that the $L_{2,1}-$norm of prediction loss item does not take into account different distances between target labels and prediction values. Another is that the $L_{2,1}-$norm of feature selection is not flexible enough to select more useful features. The third is that they usually first

[1] Data used in preparation of this article were obtained from the Alzheimer's 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.

Download English Version:

https://daneshyari.com/en/article/411463

Download Persian Version:

https://daneshyari.com/article/411463

Daneshyari.com