Neurocomputing 195 (2016) 19–22

Contents lists available at ScienceDirect

Neurocomputing

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

Graph feature selection for dementia diagnosis

Yonghua Zhu [a,b,1], Zhi Zhong [a,*], Wenfei Cao [c], Debo Cheng [d]

[a] School of Computer, Guangxi Teachers Education University, Nanning, China
[b] School of Computer, Electronics and Information, Guangxi University, Nanning, China
[c] School of Mathematics and Information Science, Shanxi Normal University, Xi'an, China
[d] Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin, China

ARTICLE INFO

Article history:
Received 7 April 2015
Received in revised form
5 September 2015
Accepted 6 September 2015
Available online 18 February 2016

Keywords:
Dementia diagnosis
Feature selection
Manifold learning

ABSTRACT

This paper proposes a graph feature selection method for dementia diagnosis, by adding the information inherent in the observations into a sparse multi-task learning framework. Specifically, this paper first defines two relations (i.e., the feature–feature relation and the sample–sample relation, respectively) based on the prior knowledge in the data. The feature–feature selection enforces the similarity relationship between features to be preserved in the coefficient matrix while the sample–sample relation is designed to preserve the relation between samples invariant in the predicted space. Then we embed these two kinds of relations into a multi-task learning framework (i.e., a least square loss function plus an l_2-norm regularization term) to conduct feature selection. Furthermore, we feed the reduced data into Support Vector Machine (SVM) for conducting the identification of Alzheimer's Disease (AD). Finally, the experimental results on a subset of the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset showed the effectiveness of the proposed method in terms of classification accuracy, by comparing with the state-of-the-art methods, including k Nearest Neighbor (kNN), ridge regression, SVM, and so on.

1. Introduction

Alzheimer's Disease (AD) is a chronic neurodegenerative disease usually starting slowly and getting worse over time since the neurons and their corresponding connections of AD patients have been demonstrated to progressively destroyed for inducing cognition function and ultimately death [1]. It has also been indicated that the identification of the early stage of AD (i.e, Mild Cognitive Impairment (MCI) – a transitional stage between normal aging and the development of dementia) enabled an earlier and more accuracy diagnosis of AD [6,7,16]. Therefore, the study on MCI has been drawn a lot of attention.

The method of neuroimaging used in helping understand the neurodegenerative process in the progression of AD study has been shown to be a powerful tool. Actually, neuroimaging methods provided a number of potential to identify individuals of AD via exploring the disease progression and therapeutic efficacy in AD and MCI. So far the main technical challenge of neuroimaging methods is to integrate effectively various baseline data for classification via concatenating the features from all kinds of baseline data to form a long vector for each subject. Machine learning methods have thus widely been designed to identify clinical labels, i.e., AD, MCI, and Normal control (NC) [2,3,9-11,14].

The AD study often encounters for the issue of high-dimensional and small-sized data due to the representation of subjects with long feature vectors. Obviously, feature selection methods (such as t-test and sparse learning methods) are very good solutions [23]. For example, Ye et al. applied sparse logistic regression for feature selection by selecting a small subset of features using the l_1-norm regularization term [8]. However, the previous methods (such as ridge regression methods and sparse feature selection methods) only make use of the correlation between samples and labels (i.e., the least square loss function) but do not utilize the information inherent in the observations. Actually, the information inherent relation has been demonstrated to play an important role as the relation between samples and labels. For example, manifold learning methods considering to preserve the information inherent in the observations outperform ridge regression in many kinds of real applications, while Chen et al. designed a graph Laplacian regularization on features captured the correlation clues to refine concept classifier, especially in cases with insufficient training samples for the application of image annotation [15].

However, conventional manifold learning methods only considered one kind of relation inherent in observation, i.e., preserving the local

[*] Corresponding author.
E-mail address: 1951488032@qq.com (Z. Zhong).
[1] Yonghua Zhu is a student at school of Computer, Electronics and Information, Guangxi University, Naning, China. He finished this draft at Guangxi Teachers' Education University where he was a visiting student supervised by the corresponding author.

http://dx.doi.org/10.1016/j.neucom.2015.09.126
0925-2312/© 2016 Published by Elsevier B.V.

structures among samples, which cannot comprehensively reflect the data distribution or the information inherent in the data [4,5].

In this paper, we put forward a new feature selection that can reflect the relationship of the dataset's inherent information to select useful features for the AD diagnosis. In proposed method, we first employ a least square loss function to construct regression for achieving the minimum regression error, and then use an l_2-norm regularization term into the regression framework for conducting feature selection. For the best use of information inherent the observations, we design two kinds of relations to construct two regularization terms, i.e., the feature–feature relation and the sample–sample relation, respectively. The sample–sample relation is designed to preserve the similarity relationship between samples, while the feature–feature relation preserves the similarity relationship between features. By integrating these two regularization terms into the sparse feature selection framework (i.e., a least square loss function plus l_2-norm regularization term), the proposed feature selection helps to select useful and important features.

The remainder of this paper is organized as follows: Section 2 describes the proposed algorithm and the novel optimization algorithm, while Section 3 analyzes the experimental results, followed by conclusions in Section 4.

2. Method

2.1. Notation

Throughout the paper, matrices are written as boldface capital letters, vectors are denoted as boldface lowercase letters, and scalars are written as normal italic letters. For a matrix $\mathbf{X} = [x_{ij}]$, its ith row, jth column are denoted as $\mathbf{x}^i$and $\mathbf{x}_j$, respectively. $\|\mathbf{X}\|_F$ is the Frobenius norm of matrix $\mathbf{X}$, and $\|\mathbf{X}\|_F = \sqrt{\sum_i \|\mathbf{x}^i\|_2^2}$. We further denote the transpose operator, the trace operator and the inverse of a matrix $\mathbf{X}$ as $\mathbf{X}^T$, $tr(\mathbf{X})$ and $\mathbf{X}^{-1}$, respectively.

2.2. Approach

In this section, we sequentially introduce the least square loss function, the sample–sample relation, and the feature–feature relation. We then proposed a new optimization method to solve the proposed objective function.

First, we use a least square loss function to map $\mathbf{X}$ into the space of $\mathbf{Y}$, by enforcing to the difference between the prediction value $\mathbf{XW}$ and clinical label $\mathbf{Y}$ as close as possible. That means to minimize the following:

$$\sum_{i=1}^{n} \sum_{j=1}^{c} (y_{ij} - \mathbf{x}^i \mathbf{w}_j)^2 = \|\mathbf{Y} - \mathbf{XW}\|_F^2 \tag{1}$$

where $\mathbf{W}$ ($\mathbf{W} \in R^{d \times c}$) is the weight coefficient and $\mathbf{x}^i$ is the ith sample of the sample matrix $\mathbf{X}$. We easily obtain the closed-form solution of Eq. (1) as $\mathbf{W} = (\mathbf{XX}^{\mathrm{T}})^{-1}\mathrm{XY}$. Due to the singular issue of $\mathbf{XX}^{\mathrm{T}}$ in real applications, a regularization term is often added to avoid this and the issue of over-fitting [4,5]. In this paper, we aim at conducting feature selection, so we use an l_2-norm regularization term (rather than using an $l_{2,1}$-norm regularization term in sparse feature selection which has been shown to generate the sparsity in a whole row [11,14]). Thus the objective function for conducting feature selection can be reconstructed as follows:

$$\min_W \|\mathbf{Y} - \mathbf{XW}\|_F^2 + \lambda \|\ \mathbf{W}\ \|_F^2 \tag{2}$$

where λ is a turning parameter and is used to adjust the sparsity of $\mathbf{W}$. However, Eq. (2) does not take any prior knowledge into account for conducting feature selection [17]. To do this, this paper defines two regularization term into Eq. (2) objective function:

$$\min_W \|\mathbf{Y} - \mathbf{XW}\|_F^2 + a_1 R_1(\mathbf{W}) + a_2 R_2(\mathbf{W}) + \lambda \|\mathbf{W}\|_F^2 \tag{3}$$

where a_1, a_2, λ are tuning parameters for balancing the magnitude among the feature–feature relation, the sample–sample relation, and the l_2-norm regularization term, respectively.

We explore the similarity relationships of any two samples and any two features. Suppose that we are given the similarity matrix $\mathbf{S} \in R^{n \times n}$ that stores the pairwise similarity scores between samples. Moreover, the larger the is s_{ij}, the more similar the ith sample and the jth sample are, and vice versa [24]. To do this, we propose to minimize the following regularization term:

$$\frac{1}{2} \sum_{i,j}^{n} s_{ij} \|\hat{\mathbf{y}}^i - \hat{\mathbf{y}}^j\|_2^2 \tag{4}$$

where s_{ij} is a weight coefficient that reflects the sample similarity between samples, $\hat{\mathbf{y}}^i$ (where $\hat{\mathbf{y}}^i = \mathbf{x}^i \mathbf{W}$) is the prediction of ith sample $\mathbf{x}^i$. The intuition behind the regularization term in Eq. (4) is that closely samples should have similar regression weights. In other words, similar samples in the original space should have similar predictions [24]. With simple mathematical transformation, we convert Eq. (4) into (5):

$$\begin{aligned} &\frac{1}{2} \sum_{i,j}^{n} s_{ij} \|\hat{\mathbf{y}}^i - \hat{\mathbf{y}}^j\|_2^2 \\ &= \frac{1}{2} \sum_{i,j}^{n} s_{ij} \|x^i \mathbf{W} - x^j \mathbf{W}\|_2^2 \\ &= \sum_{i}^{n} \mathbf{W} x^i D_{ii} (x^i)^T \mathbf{W}^T - \sum_{i,j}^{n} \mathbf{W} x^i S_{ij} (x^i)^T \mathbf{W}^T \\ &= tr\left(\mathbf{XW}(\mathbf{D} - \mathbf{S})(\mathbf{XW})^T\right) \\ &= tr\left(\mathbf{W}^T \mathbf{X}^T \mathbf{LXW}\right) \end{aligned} \tag{5}$$

where $\mathbf{D}$ is a diagonal matrix whose element is the sum of a row of $\mathbf{S}$, i.e., $D_{ii} = \sum_j S^i$. $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is the Laplacian matrix. Moreover, matrix $\mathbf{D}$ provides a natural measure on the samples, i.e., the bigger the value D_{ii} (corresponding to $\hat{\mathbf{y}}^i$) is, the more important the$\hat{\mathbf{y}}^i$ is.

Generally speaking, the similarity matrix $\mathbf{S}$ can be defined based on any reasonable measurements, such as Google distance and k Nearest Neighbors (kNN) graph. In this paper, we construct the similarity matrix $\mathbf{S}$ (or $\mathbf{M}$ in the following part of this paper) by following the previous literatures [3,9,22] with the steps: 1) constructing the adjacency graph via the kNN method, $k=3$ in our experiments according to the previous literature [12,18]. In the adjacency graph, each sample is regarded as a node and there is an edge between node i and node j if $\mathbf{x}_i$ is one of kNN of $\mathbf{x}_j$. 2) Defining the similarity for each edge, i.e., calculating the distance between two nodes [21]. To do this, this paper employs a heat kernel to define the distance between node $\mathbf{a}$ and node $\mathbf{b}$, i.e.,

$$f(\mathbf{a}, \mathbf{b}) = exp\left(-\frac{\|\mathbf{a} - \mathbf{b}\|_2^2}{2\sigma^2}\right) \tag{6}$$

where σ is the kernel width and set as 1 in our experiments according to the literatures [12].

According to the definition of the sample–sample relation in Eq. (5), we easily observe that, in Euclidean space, similar features are with small distance, i.e., large similarity. We expect that such similarity may be preserved in the new space. To this end, we define the formula of the feature–feature relation as follows:

$$\frac{1}{2} \sum_{i,j}^{d} m_{ij} \|\mathbf{w}^i - \mathbf{w}^j\|_2^2 \tag{7}$$

where m_{ij} ($m_{ij} \in \mathbf{M}^{n \times n}$) is a weight coefficient that reflects the features relation between the features [13]. By simple mathematical