

基于图的特征选择算法在阿兹海默症诊断问题研究*

朱永华¹,程德波²,何威²,文国秋²,梁正友¹

(1. 广西大学 计算机与电子信息学院, 南宁 530004; 2. 广西师范大学 计算机科学与信息工程学院, 广西 桂林 541004)

摘要: 针对神经性疾病难以确诊的问题,提出了一种基于图的特征选择方法,过滤掉不相干的特征,从而方便并且准确地对疾病患者进行诊断。算法首先基于先验知识定义了两种基本关系(特征关系和样本关系);然后将这两种关系嵌入到一个由最小二次损失函数和 l_2 -范数正则化因子组成的多任务学习框架中进行特征选择;最后,将约简得到的降维矩阵送入支持向量机(SVM)中对阿兹海默症患者进行确诊。通过对 Alzheimer's disease neuroimaging initiative(ANDI)的研究数据集进行实验得知,提出算法的分类效果均优于一般常用分类算法,如K最近邻法(KNN)、支持向量机(SVM)等。提出的算法通过考虑特征选择和引入两种数据的内在关系,有效提高了阿兹海默疾病诊断的正确率。

关键词: 阿兹海默病诊断; 特征选择; 流型学习

中图分类号: TP181 文献标志码: A 文章编号: 1001-3695(2017)04-1018-04

doi: 10.3969/j.issn.1001-3695.2017.04.014

Graph feature selection for Alzheimer's disease diagnosis

Zhu Yonghua¹, Cheng Debo², He Wei², Wen Guoqi², Liang Zhengyou¹

(1. School of Computer, Electronics & Information, Guangxi University, Nanning 530004, China; 2. College of Computer Science & Information Technology, Guangxi Normal University, Guilin Guangxi 541004, China)

Abstract: This paper proposed a graph feature selection method for Alzheimer's disease diagnosis, by adding the information inherent in the observations into a sparse multi-task learning framework. Specifically, this paper firstly defined two relations (i. e., the feature-feature relation and the sample-sample relation, respectively) based on the prior knowledge. Then embedding 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, it fed the reduced data into a 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 show 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), SVM, and so on. This paper proves that the proposed method can improve the performance of AD diagnosis accuracy, by take feature selection and two kinds of relation into account simultaneously.

Key words: Alzheimer's disease diagnosis; feature selection; manifold learning

0 引言

阿兹海默症(AD)是一种精神性疾病,潜伏期很长,随着时间的推移情况会慢慢恶化,患者神经元被逐步破坏,其认知能力也逐步下降,最终导致死亡^[1]。医学上已经证实,在患病早期,病人会出现轻度认知障碍的情况,它介于正常状态和患病状态之间,称之为MCI(mild cognitive impairment)。现在对阿兹海默症仍没有有效的治疗手段。目前研究寄希望于在患者处于MCI阶段时就能及时发现,然后采取有效措施防止病情的进一步恶化^[2]。因此,如何判别一个患者处于何种阶段成了研究阶段的重点。

利用神经影像学的方法对神经退行性疾病的诊断是十分

有效的。事实上,利用AD和MCI患者的疾病进展和治疗效果之间的关联,通过神经影像学方法可以有效预测患者到底处于何种阶段。而神经影像学的主要难题在于如何关联多个线性相关的特征,从而形成特征较少的矩阵,接着对特征较少的数据进行分析然后得出AD患者的诊断结果。基于机器学习的方法已经被成功用于疾病诊断,本文考虑对阿兹海默症患者(即AD、MCI和正常状况NC(normal control))利用机器学习进行分类诊断^[3-6]。

AD问题的研究数据通常由高维特征表示,但是由于隐私等原因能使用的病例十分少,所以常出现高特征低样本的情况。为此,特征选择的方法十分适合处理AD分类问题。例如, Ye等人^[7]利用 l_1 -范数正则式因子进行特征选择获得了一

收稿日期: 2016-03-07; 修回日期: 2016-05-06 基金项目: 国家“973”计划资助项目(2013CB329404); 中国博士后科学基金资助项目(2015M570837); 广西自然科学基金资助项目(2012GXNSFGA060004, 2015GXNSFCB139011); 国家自然科学基金资助项目(61450001, 61263035, 61573270)

作者简介: 朱永华(1994-),男,广西桂林人,本科生,主要研究方向为数据挖掘(1293234987@qq.com); 程德波(1990-),硕士研究生,主要研究方向为数据挖掘、机器学习; 何威(1989-),硕士研究生,主要研究方向为数据挖掘、机器学习; 文国秋(1987-),硕士,主要研究方向为数据挖掘、机器学习; 梁正友(1968-),教授,主要研究方向为网格计算、并行分布式计算、计算机网络。

个特征较少的子空间。然而,常用的方法(例如岭回归方法和稀疏特征选择方法)通常只利用了样本和类标签之间的关系,没有充分利用数据间的固有关系。事实上,样本和特征间的内在关联对于类别的判断起着非常重要的作用。例如,流型学习方法就是在岭回归方法的基础上,加入了数据间的固有关系,使得分类效率大大提高。因此,流型学习方法更常用于实际应用中。而利用特征间的关联性设计的图拉普拉斯正则化来改进传统的分类器,这种分类器在训练样本不足的情况下进行图像识别的效果尤为突出,这也验证了提出算法的可行性。

常用的流型学习方法通常只利用了一种数据间的关系,即只保留了样本间的结构,并没有利用数据分布情况和内在联系^[8,9]。本文提出了一种能反映数据间内在关联的特征选择方法来进行 AD 分类。该方法首先利用最小二次损失函数获取最小的回归误差,利用 l_2 -范数正则式因子进行特征选择。为了更好地利用数据间的固有关系,本文定义两种关系来获得正则化因子(即特征关系和样本关系),其中样本关系用来保持样本之间的相似性,特征关系用来保持特征间的关联性。然后,利用 SVM 对提取后的特征进行分类验证。最后,提出了一种新的优化方法,即通过在稀疏特征选择框架(即最小二次损失函数加上 l_2 -范数正则式因子)中对这两种正则式因子进行循环迭代,直至目标函数收敛,这样便可以过滤掉对特征选择带来干扰的不相干特征。

1 算法

1.1 符号注释

本文矩阵 X 的第 i 行和第 j 列分别用符号表示为 x^i 和 x_j , 矩阵 X 的范数表示为 $\|X\|_F$ 且 $\|X\|_F = \sqrt{\sum_i \|x^i\|_2^2}$ 。

1.2 算法

本节单独介绍最小二次损失函数、样本关系和特征关系。然后利用新提出的优化算法对得到的目标函数进行优化求解,本文中, X 为样本空间, Y 为类标签。

首先,利用最小二次损失函数将 X 空间降维映射到 Y 空间,使得预测值 XW 尽可能接近目标值 Y ,即确保回归误差最小化。公式表示为

$$\min \sum_{i=1}^n \sum_{j=1}^c (y_{ij} - x^i w_j)^2 = \min \|Y - XW\|_F^2 \quad (1)$$

其中: $W \in R^{d \times 1}$ 代表权重系数; x^i 表示矩阵 X 的第 i 个样本。通过线性回归方法很容易可以得到式(1)的结果,即 $W = (X^T X)^{-1} X^T Y$ 。由于 $X^T X$ 可能不可逆,使结果无意义,所以需增加一个正则化因子来避免这种情况^[8,9]。同时希望能够进行特征选择,故增加了一个 l_2 -范数正则式因子。因此,带有特征选择功能的目标函数可以写成

$$\min_W \|Y - XW\|_F^2 + \lambda \|W\|_2^2 \quad (2)$$

其中: λ 为 l_2 -范数正则式因子的参数,可以调节矩阵 W 的稀疏性。式(2)中还未加入先验知识进行特征选择。在此首先定义两种规则式,然后将其添加到目标函数式(2)中,即

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

其中: a_1, a_2, λ 为调优参数,用来分别调节式子各部分(特征关系、样本关系以及 l_2 -范数正则式因子)在式子中起到的作用,以获得更好的拟合效果。

接下来探讨任意两个样本或者任意两个特征之间的相似关系。首先,假设给定一个用来存放样本间关系系数的相似性

矩阵 $S \in R^{n \times n}$ 。在这个相似性矩阵中, s_{ij} ($s_{ij} \in S$) 越大,则代表着样本 i 和样本 j 越相似,反之亦然。为了定义样本间的相似性,最小化下面的规则式项:

$$\frac{1}{2} \sum_{i,j} s_{ij} \|\hat{y}^i - \hat{y}^j\|_2^2 \quad (4)$$

其中: S ($s_{ij} \in S$) 是一个反映样本间相似性的相似性矩阵; \hat{y}^i ($\hat{y}^i = x^i W$) 是第 i 个样本 x^i 的预测值。从式(4)可以看出,相似的两个样本有着相似的回归权重。换句话说,原始空间中两个相似的样本应该有着相同的预测值。经过简单的公式变换,可以将式(4)转换成式(5),即

$$\begin{aligned} & \frac{1}{2} \sum_{i,j} s_{ij} \|\hat{y}^i - \hat{y}^j\|_2^2 = \\ & \frac{1}{2} \sum_{i,j} s_{ij} \|x^i W - x^j W\|_2^2 = \\ & \sum_i W x^i D_{ii} (x^i)^T W^T - \sum_{i,j} W x^i S_{ij} (x^j)^T W^T = \\ & \text{tr}(XW(D - S)(XW)^T) = \\ & \text{tr}(W^T X^T L X W) \end{aligned} \quad (5)$$

其中: D 为对角矩阵,且它的对角元素是矩阵 S 中各行元素之和,即 $D_{ii} = \sum_j S_{ij}$; $L = D - S$ 为拉普拉斯矩阵。矩阵 D 衡量着该样本预测值的准确性,即 D_{ii} (对应预测值 \hat{y}^i) 越大,预测的 \hat{y}^i 值正确性就越大。

通常来说,相似性矩阵 S 可以根据不同的规则定义,例如,可以用 Google 距离,也可以利用 K 最近邻图。本文借鉴文献[4,10]来构造相似性矩阵,即利用 KNN 方法构造的邻接图来构造相似性矩阵 S (接下来小节中的矩阵 M 也由该方式构造)。而且本文根据文献[11,12]将实验参数 k 设置为 3。在邻接图中,每一个样本可以用一个点表示,如果点 x_i 是点 x_j 的 k 个最近邻点中的一个,那么,点 x_i 和点 x_j 就可以连成一条边,而这两点之间的距离(即边长)就定义为两个样本的相似程度(s_{ij})^[13]。本文利用热核来定义点 a 和点 b 之间的距离,即

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

其中: σ 是热核宽度。

根据式(5)中样本关系的定义,通过观察可以得出。在欧氏距离空间中,具有关联性的特征往往距离较近,即关联性较高。为此希望能在新空间中保持原始空间特征间的相似性,所以用下列式子表示特征关系:

$$\frac{1}{2} \sum_{i,j} m_{ij} \|w^i - w^j\|_2^2 \quad (7)$$

其中: m_{ij} ($m_{ij} \in M^{n \times n}$) 为可以反映特征间关系的相似系数矩阵。经过简单的数学变换,可以将式(7)转换成式(8):

$$\begin{aligned} & \frac{1}{2} \sum_{i,j} m_{ij} \|w^i - w^j\|_2^2 = \\ & \sum_i w^i P_{ii} (w^i)^T - \sum_{i,j} w^i m_{ij} (w^j)^T = \\ & \text{tr}(W^T (P - M) W) = \\ & \text{tr}(W^T H W) \end{aligned} \quad (8)$$

其中: P 为对角矩阵,它的每个元素对应着矩阵 M 各行元素之和,即 $P_{ii} = \sum_j m_{ij}$; $H = P - M$ 为拉普拉斯矩阵;此外,矩阵 P 中元素为对应特征重要性的量度,即 P_{ii} (对应 x^i) 越大,特征 x^i 越重要。

至此得到了如下目标函数:

$$\min_W \|Y - XW\|_F^2 + a_1 \text{tr}(W^T X^T L X W) + a_2 \text{tr}(W^T H W) + \lambda \|W\|_F^2 \quad (9)$$

式(9)中利用了两种数据间的内在联系来反映矩阵 X 的

内在关联,并将这种关系保存在系数矩阵 W 中。不同于文献 [11, 14] 中提到的方法,即在新空间中只保存原始空间距离较近的样本。提出的方法则更注重寻找样本之间的联系,并且用最小二次回归方程将样本数据 X 和类标签 Y 联系起来。

在式 (9) 中,使用了正规的表达式形式,接下来,将其转换为用迹运算表达式形式,即

$$\begin{aligned} & \text{tr}(Y - XW)^T(Y - XW) + a_1 \text{tr}(W^T X^T L X W) + \\ & a_2 \text{tr}(W^T H W) + \lambda \|W\|_F^2 \Leftrightarrow \\ & \text{tr}(Y^T - W^T X^T)(Y - XW) + \\ & a_1 \text{tr}(W^T X^T L X W) + a_2 \text{tr}(W^T H W) + \lambda \text{tr}(W^T W) \end{aligned} \quad (10)$$

化简方程后得到

$$X^T Y - X^T X W + a_1 X^T L X W + a_2 H W + \lambda W \quad (11)$$

由于式 (11) 为凸函数,所以可以运用求导的方法求出矩阵 W ,令式 (11) 等于 0,即

$$X^T Y - X^T X W + a_1 X^T L X W + a_2 H W + \lambda W = 0 \quad (12)$$

这样可以解出 W 值为

$$W = (X^T X + a_1 X^T L X + a_2 H + \lambda I)^{-1} X^T Y \quad (13)$$

其中 I 是单位矩阵。

需要注意的是,文中没有使用 l_{21} -范数正则式因子进行特征选择,而是使用 l_2 范数正则式因子生成稀疏的权重系数矩阵。然后通过寻找其中较大值保留所对应的特征,值较小所对应的特征则删除,以达到特征选择的目的。用该方法有以下两种好处:

a) 提出的方法可以获得相同的效果。由于传统的稀疏特征选择方法(即使用 l_{21} -范数正则式因子)往往通过循环迭代的方法(数十次甚至上千次)获得最后结果。所以,传统特征选择方法需要较大的时间成本,而使用 l_2 -范数正则式因子可以有效地降低时间复杂度^[15, 16]。

b) 众多文献表明,在特征选择方法中,使用 l_2 -范数正则式因子比使用 l_{21} -范数正则式因子的鲁棒性更强。因此实验中选择了 l_2 -范数正则式因子。

以下对本文方法的合理性进行简单的理论分析。根据式 (9) 可知,提出的算法主要由三大理论框架组成: a) l_2 正则式范数,即 $\|W\|_F^2$; b) 图拉普拉斯算子,即 $\text{tr}(W^T X^T L X W)$ 和 $\text{tr}(W^T H W)$; c) 最小二乘损失函数(least square loss function),即 $\min \|Y - XW\|_F^2$ 。

a) 利用替换正则式范数, Zhang 等人在研究中从理论和实验均证明了正则式范数具有高效的运算速度、显著的分类性能和较强的鲁棒性^[17]。

b) 式 (4) 表明计算样本预测值 \hat{y}^i, \hat{y}^j 时,将它们原始空间的相似性嵌入到相似性矩阵 S ,而根据文献 [1, 11] 可知,若特征空间中的任意两个样本 x^i, x^j 具有相似性,则它们预测的结果 \hat{y}^i, \hat{y}^j 也必然相似。同理式 (7) 也具有同样的性质^[11]。

c) 在阿兹海默患者数据采集过程中,由于其涉及到隐私问题,往往只有小样本数据集,而此类数据集通常由大量特征表示。这使得在小样本高维数据上的建模极具挑战性。对该类数据集进行分类处理时,即使只对同类样本进行训练,最后得到的拟合误差 $(Y - XW)$ 也会较大,并最终造成分类性能不稳定。解决此问题的方法除了十折交叉验证之外,另一种有效方法即使用最小二乘损失函数。假设 $X = [X_1, X_2, \dots, X_k]$ 其中 k 为数据集的类别数, X_k 为第 k 类样本。最小误差损失函数归结起来就是计算 $W = \text{argmin} \|Y - XW\|$ 的问题。此时,假设最小拟合误差为: $e = \|Y - XW\|_2^2 = \|y - \hat{y}\|_2^2 + \|\hat{y} - xw\|_2^2$ 。其中在不同

类别中,式子 $\|y - \hat{y}\|_2^2$ 均为固定值,所以最终决定拟合误差的大小在于预测值 $\hat{y} = \sum_1^k x_i w_i$,其几何表示如图 1 所示。

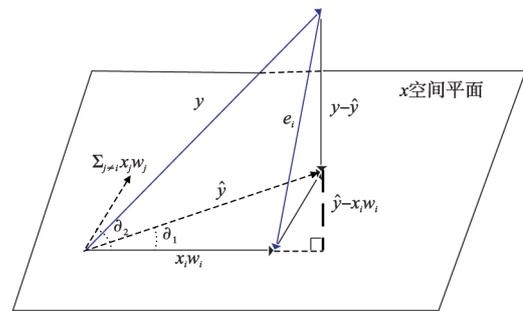


图1 最小误差损失式的空间表示

最小误差原则需要使重构或分类误差保持最小,即 $e^* = \|\hat{y} - xw\|_2^2$ 保持最小,由图 1 可知 $\sum_{j \neq i} x_j w_j$ 与 $\hat{y} - x_i w_i$ 平行,而它们具有相同的垂线,即图 1 中粗线,此时易有如下等式: $\frac{\|\hat{y}\|_2}{\sin(\theta_1)} = \frac{\|\hat{y} - x_i w_i\|_2}{\sin(\theta_2)}$ 其中 θ_1 和 θ_2 分别为夹角 $(\hat{y}, x_i w_i)$ 和

角 $(\sum_{j \neq i} x_j w_j, x_i w_i)$ 。上述公式推导得: $e^* = \frac{\sin(\theta_2) \times \|\hat{y}\|_2}{\sin(\theta_1)}$ 。

由此,在通过特征选择进行分类的过程,预测值 \hat{y} 若想满足最小误差的要求,需要至少同时满足以下两条件,即 θ_2 尽可能小和 θ_1 尽可能大。如此则双重限制使得重构误差的有效性和鲁棒性得到了理论上的保证。

1.3 特征选择

通过解式 (9),可以得到一个权重系数矩阵 W 。其中, W 的每一行都对应着 X 空间的一个特征。将 W 按行求和然后按值的大小进行排列,保存值较大元素所对应的特征,也就是说这些特征对结果的判断是重要不可或缺的。实验还使用了十折交叉验证避免实验的随机性导致的误差。通过这种方法就能筛选出不相干的冗余特征,获得一个降维后的矩阵,然后将该降维矩阵送入支持向量机(SVM)中获得分类结果。

2 实验分析

本文利用提出算法和对比方法分别对三组二进制数据集进行了实验。

2.1 实验设置

从 ADNI 数据集中选择三个子数据集进行二进制分类实验,它们分别是 ADs vs NCs(生病状态和正常状态)、MCIs vs NCs(中间状态和正常状态)和 MCI-Converters (MCI-Cs) vs MCI-Non-Converts (MCI-NCs)(中间状态和非中间状态,即正常状态和生病状态)。

由于数据集的特征多而样本数量相对较少,很难构造一个有效的模型对患者进行疾病诊断。所以,在实验中运用了十折交叉验证的方法。即将数据集的样本平均分成十份,其中一份做测试集,剩余九份做训练集。在处理训练集样本中,又使用了五折交叉验证进行特征筛选。选出分类结果最好的参数作为测试集的参数。实验重复十次,取十次结果的平均值作为实验最终结果。实验参数设置如下: $\lambda \in (10^0, 10^1, \dots, 10^6)$ $a_1 \in (10^{-3}, 10^{-2}, \dots, 10^4)$ $a_2 \in (10^{-3}, 10^{-2}, \dots, 10^4)$ $k=3$ $\sigma=1$ 。

2.2 对比实验方法

为了证明提出算法的分类效果,常用的线性方法(即不进

行特征选择方法,如 KNN、SVM 和 RR) 和最近流行的方法(具有特征选择的方法,如 MFS) 分别对数据集进行了对比实验。对比方法主要细节陈述如下:

- a) KNN。K 最近邻算法通过对 k 设定不同参数,用十折交叉验证方法获得最佳分类结果。
- b) SVM。支持向量机通过设定不同的内核参数,然后用十折交叉验证来获得最佳分类结果。
- c) RR。岭回归(ridge regression) 方法通过对所有的特征进行分析获得分类结果,即不进行特征选择。
- d) MFS。流型学习的特征选择方法(manifold feature selection) 进行特征选择,即只运用了样本关系,根据对式(10)中设定的特定参数实现。

2.3 实验结果

表 1 展示了所有实验的分类结果,值越高则意味着该分类方法效果越好。通过这些数据,可以获得以下几点:

- a) 提出算法取得了最好的分类结果,接下来分别是 MFS、SVM、RR 和 KNN。其中,提出算法对比比实验中获得最好结果的 MFS 方法分别提高了 0.13%、1.51% 和 0.23%,对比比实验中获得最差结果的 KNN 方法提高了 19.46%、12.96% 和 15.89%(实验结果分别对应数据集 AD vs. NC, MCI vs. NC 和 MCI-C vs. MCI-C) 特征。

表 1 所有实验数据的分类情况(平均值)

算法	AD vs. NC	MCI vs. NC	MCI-C vs. MCI-NC
KNN	71.78(0.28)	61.44(0.54)	61.34(2.14)
SVM	89.0(0.86)	64.28(0.35)	64.86(1.65)
RR	80.6(1.27)	67.62(0.34)	65.83(1.52)
MFS	91.10(0.18)	72.79(1.15)	77.00(1.85)
本文	91.24(0.28)	74.40(0.84)	77.23(2.08)

注:黑体字为最佳结果,括号内为多次实验结果方差。

特征选择方法(MFS 和提出方法)比非特征选择方法(RR、SVM 和 KNN) 效果更好,分别提高了 8.91%、9.27% 和 12.50%(实验结果分别对应数据集 AD vs. NC, MCI vs. NC 和 MCI-C vs. MCI-NC)。由此可以看出,在阿兹海默症的诊断上来看,进行特征选择可以有效且大幅地提高疾病诊断的准确率。

- b) 对比另一个特征选择方法 MFS,提出算法的分类效果更好,分别提高了 0.14%、1.61%、0.63%(实验结果分别对应数据集 AD vs. NC, MCI vs. NC 和 MCI-C vs. MCI-NC),由于提出算法利用了更多数据间的内在关系,即在原有的 MFS 方法上添加了样本特征关系,在稀疏过程中更好地保存了数据的分布结构,所以该方法的分类效果更好。

3 结束语

通过在稀疏性特征选择的框架中嵌入数据间的固有关系,本文提出了一种基于图的特征选择方法,即在由一个最小二次损失函数加上 l_2 -范数正则化因子组成的多任务学习框架中嵌入特征关系和样本关系,并利用该模型与 SUM 结合对阿兹海默症患者进行诊断。对 ADNI 的三个数据集进行了多次实验,结果均表明:在处理多特征、低样本的数据集时,进行特征选择和使用数据间的固有联系均能有效地提高样本分类的准确率。

参考文献:

[1] Zhu Xiaofeng, Suk H, Shen Dinggang. Matrix-similarity based loss

function and feature selection for Alzheimer's disease diagnosis [C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Press, 2014: 3089-3096.

- [2] Zhu Xiaofeng, Suk H, Shen Dinggang. A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis[J]. *NeuroImage* 2014, 100: 91-105.
- [3] Cheng Bo, Zhang Daoqiang, Chen Songcan, et al. Semi-supervised multimodal relevance vector regression improves cognitive performance estimation from imaging and biological biomarkers[J]. *Neuroinformatics* 2013, 11(3): 339-353.
- [4] Franke K, Ziegler G, Klöppel S, et al. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters[J]. *NeuroImage* 2010, 50(3): 883-892.
- [5] Seppälä T T, Koivisto A M, Hartikainen P, et al. Longitudinal changes of CSF biomarkers in Alzheimer's disease[J]. *Journal of Alzheimer's Disease* 2011, 25(4): 583-594.
- [6] Weinberger K Q, Sha Fei, Saul L K. Learning a kernel matrix for nonlinear dimensionality reduction[C]//Proc of the 21st International Conference on Machine Learning. New York: ACM Press, 2004: 106.
- [7] Ye Jieping, Farnum M, Yang E, et al. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data[J]. *BMC Neurology* 2012, 12(1): 1-12.
- [8] Golub G H, Van Loan C F. Matrix computations[M]. 3rd ed. [S. l.]: Johns Hopkins University Press, 1996.
- [9] Nie Feiping, Huang Heng, Cai Xiao, et al. Efficient and robust feature selection via joint l_2, l_1 -norms minimization[C]//Advances in Neural Information Processing Systems. 2010: 1813-1821.
- [10] Zhu Xiaofeng, Huang Zi, Cui Jiangtao, et al. Video-to-shot tag propagation by graph sparse group Lasso[J]. *IEEE Trans on Multimedia*, 2013, 15(3): 633-646.
- [11] Belkin M, Niyogi P, Sindhvani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled example[J]. *Journal of Machine Learning Research* 2006, 7(1): 2399-2434.
- [12] Zhu Xiaofeng, Zhang S, Jin Zhi, et al. Missing value estimation for mixed-attribute data sets[J]. *IEEE Trans on Knowledge & Data Engineering* 2011, 23(1): 110-121.
- [13] Zhu Xiaofeng, Huang Zi, Shen Hengtao, et al. Dimensionality reduction by mixed kernel canonical correlation analysis[J]. *Pattern Recognition* 2012, 45(8): 3003-3016.
- [14] Zhu Xiaofeng, Huang Zi, Yang Yang, et al. Self-taught dimensionality reduction on the high-dimensional small-sized data[J]. *Pattern Recognition* 2013, 46(1): 215-229.
- [15] Zhu Xiaofeng, Huang Zi, Shen Hengtao, et al. Linear cross-modal hashing for efficient multimedia search[C]//Proc of the 21st ACM International Conference on Multimedia. New York: ACM Press, 2013: 143-152.
- [16] Zhu Xiaofeng, Huang Zi, Cheng Hong, et al. Sparse hashing for fast multimedia search[J]. *ACM Trans on Information Systems* 2013, 31(2): 9.1-9.24.
- [17] Liang Lei, Yang Meng, Feng Xiangchu. Sparse representation or collaborative representation: which helps face recognition[C]//Proc of International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2011: 471-478.