软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(12):4545−4558 [doi: 10.13328/j.cnki.jos.006376]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于标签对齐的多模态一致性表型关联方法*

汪美玲[1,2], 邵 伟[1,2], 张道强[1,2]

[1](南京航空航天大学 计算机科学与技术学院, 江苏 南京 211106)

[2](模式分析与机器智能工业和信息化部重点实验室, 江苏 南京 211106)

通信作者: 张道强, E-mail: dqzhang@nuaa.edu.cn

摘 要: 近年来, 随着脑影像和基因技术的发展, 脑影像遗传学得到了广泛的关注. 在脑影像遗传研究中, 检验遗传变异(即单核苷酸多态性(single nucleotide polymorphisms, SNPs))对大脑结构或功能的影响是一项艰巨的任务. 此外, 提取的多模态脑表型和来自同一区域的一致性脑影像标志物为理解疾病(例如, 阿尔茨海默病(Alzheimer's disease, AD))的机理提供了更多的见解. 利用多模态脑表型作为桥接风险基因位点和疾病状态的中间特征, 设计通过标签对齐的多模态学习方法来识别 AD 中风险基因位点与疾病状态之间的一致性表型. 首先, 用标准的多模态方法去探索和 AD 相关的基因位点(即 APOEe4 rs429358)与多模态脑影像之间关系; 其次, 为了利用标记样本之间的标签信息, 在标准多模态方法的目标函数中添加了一个新的标签对齐正则化项, 使得所有具有相同类别标签的多模态样本在映射空间中更靠近; 最后, 在公开的 ADNI (Alzheimer's disease neuroimaging initiative)数据集上的 3 种脑影像(即大脑的结构组织信息、脱氧葡萄糖正电子发射断层扫描和正电子发射断层扫描淀粉样蛋白成像)进行实验. 实验结果表明: 该方法可以在多模态脑影像上发现鲁棒的、一致性脑区域来解释 AD 的病因, 并在 3 个模态上将相关系数分别提高了 8%, 9%, 5%.

关键词: 脑影像遗传学; 多模态脑影像表型; 单核苷酸多态性; 标签对齐; 阿尔茨海默病

中图法分类号: TP18

中文引用格式: 汪美玲, 邵伟, 张道强. 基于标签对齐的多模态一致性表型关联方法. 软件学报, 2022, 33(12): 4545−4558. http://www.jos.org.cn/1000-9825/6376.htm

英文引用格式: Wang ML, Shao W, Zhang DQ. Label-aligned Multi-modality Consistent Phenotype Association Method. Ruan Jian Xue Bao/Journal of Software, 2022, 33(12): 4545−4558 (in Chinese). http://www.jos.org.cn/1000-9825/6376.htm

Label-aligned Multi-modality Consistent Phenotype Association Method

WANG Mei-Ling[1,2], SHAO Wei[1,2], ZHANG Dao-Qiang[1,2]

[1](College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

[2](MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China)

Abstract: Recently, with the rapid development of imaging and genomic techniques, the brain imaging genetics has received extensive attention. In the brain imaging genetic studies, it is a challenging task to examine the influence of genetic variants, i.e., single nucleotide polymorphisms (SNPs), on structures or functions of human brains. In addition, multimodal brain imaging phenotypes extracted from different perspectives and imaging markers from the same region consistently showing up in multimodalities gives more ways to understand the diseases mechanism, such as Alzheimer's disease (AD). Accordingly, This work exploits multi-modal brain imaging phenotypes as intermediate traits to bridge genetic risk factors and disease status. Consistent phenotype between genetic risk factors and disease status is discovered via the designed label-aligned multi-modality regression method in AD. Specifically, standard multi-modality method is first applied to explore the relationship between the well-known AD risk SNP APOEe4 rs429358 and multimodal brain imaging

* 基金项目: 国家自然科学基金(61876082, 61902183, 61861130366, 61732006); 国家重点研发计划(2018YFC2001600, 2018YFC2001602, 2018ZX10201002); 牛顿高级学者基金(NAF\R1\180371)

收稿时间: 2020-08-01; 修改时间: 2021-02-03; 采用时间: 2021-05-06

phenotypes. Secondly, to utilize the label information among labeled subjects, a new label-aligned regularization is included into the standard multi-modality method. In such way, all multimodality subjects with the same class labels should be closer in the new embedding space. Finally, the experiments are conducted on three baseline brain imaging modalities, i.e., voxel-based measures extracted from structural magnetic resonance imaging, fluorodeoxyglucose positron emission tomography and F-18 florbetapir PET scans amyloid imaging, from the Alzheimer's disease neuroimaging initiative (ADNI) database. Related experimental results validate that the proposed method can identify robust and consistent regions of interests over multi-modality imaging data to guide the disease-induced interpretation. Furthermore, the values of correlation coefficient have been increased by 8 %, 9%, and 5% in comparison with the best results of the existing algorithms on three modalities.

Key words: brain imaging genetics; multimodal brain imaging phenotypes; single nucleotide polymorphisms; label-aligned; Alzheimer's disease

随着社会的发展, 脑疾病的患病率及发病率均在不断上升. 阿尔茨海默病(Alzheimer's disease, AD)是最常见的脑疾病之一, 是一种持续性高级神经功能活动障碍, 俗称老年痴呆症, 其患病年龄段主要是在老年期以及老年前期. 研究如何精确诊断 AD, 特别是患病早期阶段的轻度认知障碍, 是当前的一个热点问题[1−8]. 在过去的几十年中, 脑影像遗传学在脑疾病研究领域引起了广泛关注[9−20], 其目标是如何找到遗传标记物(单核苷酸多态性(single nucleotide polymorphisms, SNP))[21]与多模态神经影像数据分离的表型特征(quantitative traits, QT)之间的关联. 而挖掘神经影像和基因数据与疾病之间的关联关系及演化规律、寻找与 AD 等脑疾病的生物标志物, 可为复杂疾病的发病机理提供数据依赖的解释, 为实现诊断预测甚至早期治疗提供基础支撑.

早期的脑影像遗传学研究已经进行了全基因组关联研究(genome-wide association studies, GWAS)[22,23]. 为了解决影像遗传学数据的高维问题, 研究人员提出了一些假设驱动的方法, 将重点放在少数遗传变量上, 并在整个大脑中寻找它们的关联 QT[24,25]. 还有一些研究则集中在有限数量的影像 QT 上, 并在整个基因组中搜索它们的关联 SNP[26,27]. 然而, 上述方法只考虑了风险基因和脑影像 QTs 之间的关联. 一般来说, 特定脑疾病(如 AD)会在对应风险基因所调控的脑影像 QTs 上呈现出异常. 在影像遗传学关联研究中, 希望找到可以从风险基因位点到致病脑区再到疾病特异性的关联分析模型. 有研究表明, 分析单基因位点变异对多脑区的影响可以很好地帮助我们理解脑部疾病[28−30]. 但上述工作仅仅考虑了基因与脑结构表型(structural magnetic resonance imaging, MRI)间的关联, 而且这些脑结构影像为单模态数据, 其无法较为全面地表征大脑的属性特征. 对于多模态方式, 起初是用于医学影像分析领域, 主要进行疾病诊断和预测[31−33]. 近些年, 多模态分析的方式也被用在全基因组关联研究中, 用来改善风险基因 SNP 和多脑区影像 QTs 间的关联性能, 从不同角度(例如结构和功能等)来理解致病机理[34]. 例如, Hao 等人[35]在多模态脑影像关联模型的基础上, 考虑了临床诊断结果(例如 NC, SMC, EMCI, LMCI 和 AD 等标号)作为影像遗传学分析的诱导信息, 能够辅助检测出同时与疾病和风险基因关联的那些脑区 QTs 特征. 通过引入可以进行类别相似性度量约束, 即 Laplacian 正则化项, 提出了基于诊断信息引导的多模态方法(diagnosis-guided multi-modality, DGMM). 但是, 这些方法仅关注相同样本的多模态信息, 而忽略了不同模态不同样本间的内在联系. 相同样本内的多模态信息只是通过嵌入多模态数据之间的互补信息来提高选择更具有判别性特征的能力. 而考虑不同样本间的内在联系能最大限度地将模态与模态之间的结构信息嵌入到目标函数中, 从而能够诱导出与风险基因位点具有更强关联的判别脑区, 帮助我们从结构和功能等不同角度理解致病机理.

为解决上述问题, 本文提出了利用多模态表型数据作为桥接风险基因位点和疾病状态的中间特征, 通过设计标签对齐的多模态学习方法在 AD 中发现风险基因位点与疾病状态之间的一致性表型. 本文的主要贡献如下:

(1) 为尽可能多地保持不同模态之间和模态自身的类别结构信息, 本文利用组稀疏化项, 能够确保与风险基因和疾病状态同时相关的脑区特征能被联合地从多模态数据中选择出来, 提高了判别能力;

(2) 考虑特定疾病(如 AD)的风险基因型(如 APOE e4 rs429358)与该疾病的表型 QTs 之间具备高度的相关性, 本方法采用标签对齐正则化项, 能够提取出更具判别力的脑影像特征, 作为基因型与疾病状态之间桥梁. 此外, 本文提出的标签对齐正则化项是 Laplacian 正则化项的一个推广, 因而更具普适性;

(3) 在 Alzheimer's Disease Neuroimaging Initiative (ADNI)数据库中的 3 种脑影像, 即大脑的结构组织信息(voxel-based measures extracted from structural magnetic resonance imaging, VBM-MRI)、脱氧葡萄糖正电子发射断层扫描(fluorodeoxyglucose positron emission tomography, FDG-PET)和正电子发射断层扫描淀粉样蛋白成像(F-18 florbetapir PET scans amyloid imaging, AV45-PET), 以及基因数据集上进行实验. 实验结果表明: 在这 3 种脑影像数据上, 相比于现有最好的多模态学习方法, 本文提出的方法在相关系数指标上分别提高了 8% (VBM)、9% (FDG)和 5%(AV45).

1 相关工作

假定脑影像表型为 $X=[x_1,\ldots,x_n,\ldots,x_N]^T\in R^{N\times d}$, 基因型为 $Y=[y_1,\ldots,y_n,\ldots,y_N]^T\in R^N$, 其中, N 是样本数, d 是指脑表型的特征维数. 基于此, 一个常见且基本的风险基因与脑影像关联的目标函数如下:

$$\min_w \frac{1}{2}\|Y-Xw\|_2^2+\lambda\|w\|_1 \tag{1}$$

其中, λ是正则化参数, w 中非零元素是指与回归输出相关的输入特征.

值得注意的是: 这些研究只考虑基因与脑结构表型 MRI 间的关联, 并且该脑结构影像为单模态数据, 其无法较为全面地表征大脑属性特征. 因此, 许多相关研究采用了多模态方式来改善关联性能.

1.1 多模态一致性表型关联

受文献[31–33]等工作启发, 多模态分析的方式被用在全基因组关联研究中, 用来改善风险基因 SNP 和多脑区影像 QTs 间的关联性能, 从不同角度(例如结构和功能等)来理解致病机理. 假设有 M 个模态的数据表型, 其中, 第 m 个模态的数据矩阵为 $X^m=[x_1^m,...,x_n^m,...,x_N^m]^T\in R^{N\times d}$, N 是样本数, d 对应每种模态脑表型的特征维数. 基因型 $Y=[y_1,\ldots,y_n,\ldots,y_N]^T\in R^N$ 为 N 个样本对应的响应向量. 基于此, 风险基因位点与多模态脑影像(multi-modality, MM)关联的目标函数如下:

$$\min_W \frac{1}{2}\sum_{m=1}^{M}\|Y-X^mw^m\|_2^2+\lambda\|W\|_{2,1} \tag{2}$$

其中, w^m 是第 m 个模态的权重向量; $W=[w^1,\ldots,w^m,\ldots,w^m]^T\in R^{d\times M}$ 是相应模态上的权重向量组成的权重矩阵; λ是正则化参数; $\|W\|_{2,1}=\sum_{j=1}^{d}\|w_j\|_2$ 是一个 $l_{2,1}$ 组稀疏正则化项, 用于联合地选择多模态脑影像中少数与风险基因位点相关的脑区特征. 当 M=1 时, 该模型退化为基因与单模态脑影像的关联模型(见公式(1)).

1.2 基于诊断信息引导的多模态一致性表型关联

Hao 等人[35]在多模态脑影像关联模型的基础上, 考虑了临床诊断结果(例如 NC, SMC, EMCI, LMCI 和 AD 等标号)作为影像遗传学分析的诱导信息, 能够辅助检测出同时与疾病和风险基因关联的那些脑区 QTs 特征. 通过引入可以进行类别相似性度量约束, 即 Laplacian 正则化项, 提出了基于诊断信息引导的多模态模型 DGMM:

$$\min_W \frac{1}{2}\sum_{m=1}^{M}\|Y-X^mw^m\|_2^2+\lambda_1\|W\|_{2,1}+\lambda_2\sum_{m=1}^{M}(X^mw^m)^TL^mX^mw^m \tag{3}$$

其中, w^m 是第 m 个模态的权重向量; $W=[w^1,\ldots,w^m,\ldots,w^m]^T\in R^{d\times M}$ 是相应模态上的权重向量组成的权重矩阵; λ_1 和λ_2 是正则化参数; $\|W\|_{2,1}=\sum_{j=1}^{d}\|w_j\|_2$ 是一个 $l_{2,1}$ 组稀疏正则化; $L^m=D^m-S^m$ 是对应的多模态 Laplacian 矩阵, D^m 是对角矩阵, $S^m=[s_{ij}^m]$ 是相似度矩阵, s_{ij}^m 表示在第 m 个模态上样本 x_i^m 和样本 x_j^m 的相似度, 其定义如下:

$$s_{ij}^m=\begin{cases}1, & \text{如果}x_i^m\text{和}x_j^m\text{属于同一类}\\ 0, & \text{否则}\end{cases} \tag{4}$$

公式(4)说明: 如果样本 x_i^m 和 x_j^m 来自同一个类, 则 w^Tx_i 与 w^Tx_j 之间距离就越小.

2 基于标签对齐的多模态一致性表型关联

虽然上述的风险基因位点与多模态脑影像关联模型 MM 和基于诊断信息引导的多模态模型 DGMM 能够很好地利用多模态间的互补信息, 进一步地改善特征选择的能力, 但是它们仅限于相同样本的多模态信息, 不同模态不同样本间的内在联系没有得到充分利用.

2.1 基于标签对齐的多模态一致性表型关联

基于上述考虑, 为了在关联学习过程中既充分利用同一模态内部的类别结构信息, 又充分利用不同模态之间的信息, 我们引入标签对齐正则化项:

$$\min_{W}\sum_{i,j}^{N}\sum_{p,q}^{M}\|(w^p)^T x_i^p-(w^q)^T x_j^q\|_2^2 s_{ij} \tag{5}$$

其中,

$$s_{ij}^m=\begin{cases}1, & \text{如果}x_i^p\text{和}x_j^q\text{属于同一类}\\ 0, & \text{否则}\end{cases} \tag{6}$$

公式(6)说明: 如果样本 x_i^p 和 x_j^q 来自同一类, 则 $(w^p)^T x_i^p$ 和 $(w^q)^T x_j^q$ 之间的距离就越小. 对于公式(5), 当两个模态互不相同时, 不同模态的互补信息将能够指导疾病相关的脑影像与风险基因位点的关联. 当为同一模态时, 通过标签对齐方式, 可以获得模态的几何类别结构信息, 即通过嵌入标签信息来诱导疾病相关的脑影像与风险基因位点的关联. 因此, 公式(5)既能够描述不同模态所传递的补充信息, 也可以保持多模态间的内在诊断相关性. 在公式(5)的基础之上, 我们提出了新的基于标签对齐的多模态(label-aligned multi-modality, LAMM)模型, 其数学描述如下:

$$\min_{W}\frac{1}{2}\sum_{m=1}^{M}\|Y-X^m w^m\|_2^2+\lambda_1\|W\|_{2,1}+\lambda_2\sum_{i,j}^{N}\sum_{p,q}^{M}\|(w^p)^T x_i^p-(w^q)^T x_j^q\|_2^2 s_{ij} \tag{7}$$

其中, w^m 是第 m 个模态的权重向量; $W=[w^1,\ldots,w^m,\ldots,w^m]^T\in R^{d\times M}$ 是相应模态上的权重向量组成的权重矩阵; λ_1 和 λ_2 是正则化参数; $\|W\|_{2,1}=\sum_{j=1}^{d}\|w_j\|_2$ 是一个 $l_{2,1}$ 组稀疏正则化项. 在模型(7)中: 组稀疏化项 $\|W\|_{2,1}$ 用于联合地选择多模态脑影像中少数与风险基因位点相关的脑区特征; 标签对齐正则化项是通过利用标签信息来对齐各模态的样本信息, 使得同类样本不同模态标签的相关性信息得以保持, 这样就可以诱导出与风险基因位点之间具备强关联的判别脑区, 其结构框架如图 1 所示($\|W\|_{2,1}$ 为组稀疏化项, 用于联合地选择多模态脑影像中少数与风险基因位点相关的脑区特征; $\sum_{i,j}^{N}\sum_{p,q}^{M}\|(w^p)^T x_i^p-(w^q)^T x_j^q\|_2^2 s_{ij}$ 是标签对齐正则化项, 通过利用标签信息来对齐各模态的样本信息, 使得同类样本不同模态标签的相关性信息得以保持).

另外, 当 $p=q=m$ 时, 则有:

$$\min_{W}\sum_{i,j}^{N}\sum_{p,q}^{M}\|(w^p)^T x_i^p-(w^q)^T x_j^q\|_2^2 s_{ij}=\min_{W}\sum_{i,j}^{N}\sum_{m}^{M}\|(w^m)^T x_i^m-(w^m)^T x_j^m\|_2^2 s_{ij}=2\sum_{m}^{M}(X^m w^m)^T L^m X^m w^m \tag{8}$$

其中, $L^m=D^m-S^m$ 是对应的多模态 Laplacian 矩阵, D^m 是对角矩阵, $S^m=[s_{ij}^m]$ 是相似度矩阵, s_{ij}^m 表示在第 m 个模态上样本 x_i^m 和样本 x_j^m 的相似度. 因此, 该标签对齐正则化项(5)退化为 Laplacian 正则化项. 这意味着, 基于诊断信息引导的多模态模型(DGMM)是我们提出的基于标签对齐的多模态模型(LAMM)的一个特例. 这说明我们提出的方法更具普遍性.

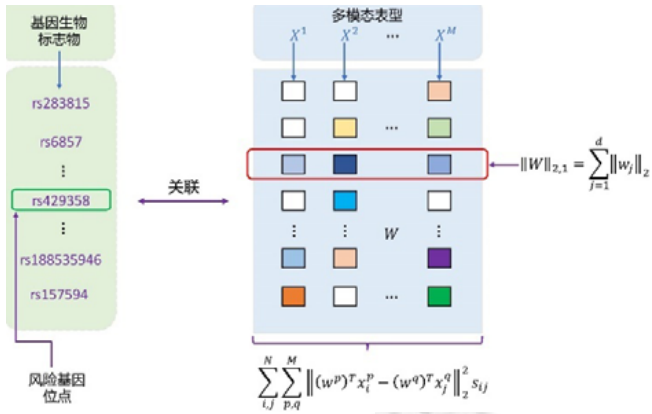


图 1　提出的 LAMM 方法在风险基因位点与多模态表型的关联过程

2.2 算法优化过程

对于模型(7), 我们可以使用加速近似梯度(accelerated proximal gradient , APG)[36]来获得其最优解. 首先, 模型(7)分成平滑子式 $f(W)$和非平滑子式 $g(W)$如下:

$$f(W)=\frac{1}{2}\sum_{m=1}^{M}\|Y-X^m w^m\|_2^2+\lambda_2\sum_{i,j}^{N}\sum_{p,q}^{M}\|(w^p)^T x_i^p-(w^q)^T x_j^q\|_2^2 s_{ij} \tag{9}$$

$$g(W)=\lambda_1\|W\|_{2,1} \tag{10}$$

然后, 定义一个近似函数$\Omega(W,W_i)$如下:

$$\Omega(W,W_i)=f(W_i)+\langle W-W_i,\nabla f(W_i)\rangle+\frac{l}{2}\|W-W_i\|_F+g(W) \tag{11}$$

其中, $\|\cdot\|_F$是指 Frobenius 范数, $\nabla f(W_i)$表示 $f(W)$第 i 次迭代在 W_i 处的梯度, l 是迭代步长. 故, APG 的更新如下:

$$W_{i+1}=\arg\min_W\frac{1}{2}\|W-V\|_F^2+\frac{1}{l}g(W)=\arg\min_{w_1,\ldots,w_d}\frac{1}{2}\sum_{j=1}^{d}\left(\|w_j-v_j\|_2^2+\frac{\lambda_1}{l}\right)\|w_j\|_2 \tag{12}$$

其中, w_j和 v_j 分别是指矩阵 W 和矩阵 V 的第 j 列, 且:

$$V=W_i-\frac{1}{l}\nabla f(W_i) \tag{13}$$

因此, 通过公式(13), 优化问题可以分解成 d 个子问题, 且这些子问题有如下解析解:

$$w_j^*=\begin{cases}\left(1-\dfrac{\lambda_1}{l\|v_j\|_2}\right)v_j, & \text{如果}\ \|v_j\|_2>\dfrac{\lambda_1}{l}\\ 0, & \text{否则}\end{cases} \tag{14}$$

此外, 在执行梯度下降时, 通过搜索点 Q_i 来代替 W_i, 其计算形式如下:

$$Q_i=W_i+\alpha_i(W_i-W_{i-1}) \tag{15}$$

且:

$$\alpha_i=\frac{(1-\rho_{i-1})\rho_i}{\rho_{i-1}} \tag{16}$$

$$\rho_i=\frac{2}{i+3} \tag{17}$$

上述优化算法可归纳为如下.

算法 1. LAMM 优化算法.

输入: 风险基因位点 APOEe4 rs429358 $Y=[y_1,\ldots,y_n,\ldots,y_N]^T\in R^N$, 多模态影像数据 $X^m=[x_1^m,\ldots,x_n^m,\ldots,x_N^m]^T\in R^{N\times d}$, 样本的诊断类别信息(NC, SMC, EMCI, LMCI 和 AD);

输出: W.

1. **FOR** i=1 to I **do**
2. 根据公式(15)计算搜索点 Q_i;
3. l_i=l_{i-1};
4. 当 $f(W_{i+1})+g(W_{i+1})>\Omega(W_{i+1},Q_i)$;
5. l_i=σl_{i-1};
6. 根据公式(12)更新 W_{i+1};

此外, 因为 f 是凸函数, 所以可以得到如下不等式:

$$\|\nabla f(W)-\nabla f(U)\|_F\leqslant c\|W-U\|_F \tag{18}$$

其中, c 为正的常数. 令 $F(W)=f(W)+g(W)$, 假设该方程的最优解为 W^*, 即:

$$F(W^*)\leqslant F(W_i) \tag{19}$$

然后, 根据文献[35], 得到如下不等式:

$$F(W_i)-F(W^*)\leqslant\frac{2\tau c\|W_0-W^*\|_F}{(i+1)^2} \tag{20}$$

因此, 根据预先设定的条件:

$$F(W_i)-F(W^*)\leqslant\varepsilon \tag{21}$$

则, 算法 1 的迭代次数为

$$\left\lfloor\sqrt{\frac{2\tau c\|W_0-W^*\|_F}{\varepsilon}}-1\right\rfloor\to \mathrm{O}\left(\frac{1}{\sqrt{\varepsilon}}\right) \tag{22}$$

因此, 算法 1 的收敛速度是 $\mathrm{O}\left(\frac{1}{l^2}\right)$, 其中, 字母 I 表示算法的最大迭代次数. 另外, 算法 1 的主要计算资源消耗来自于 $f(W)$和优化公式(12), 计算 $f(W)$梯度的时间复杂度为 O(dN), 迭代更新公式(12)的计算复杂度为 O(dM), 因此, 单次迭代所对应的时间复杂度为 O(d(M+N)). 最终可知, 算法 1 经过 I 次迭代之后的时间复杂度为 O(d(M+N)I).

3 实验结果与分析

3.1 实验设置

采用 5 折交叉验证的方法, 把整个样本集合平均划分成 5 个部分, 每次随机取 4 个部分作为训练集来进行内部交叉验证, 并利用网格搜索(grid search)来实现参数选择, 参数范围为{10^{-5},3×10^{-5},10^{-4},3×10^{-4},…,3,10}, 剩下一部分作为测试集. 此外, 使用实际和预测响应之间的相关系数评估方法的关联性能, 该指标被广泛用于衡量回归和关联分析的效果.

我们对比一些最新方法来评估我们提出方法的性能. 详细信息如下.

1) 单模态(single modality, SM), 即风险基因与单个脑影像关联[28−30];
2) 多模态(multi-modality, MM), 即风险基因与多模态脑影像关联[31−33];
3) 诊断信息引导的多模态方法(diagnosis-guided multi-modality, DGMM)[35];
4) 加入本文提出的标签对齐信息的 SM(即基于标签对齐的单模态(label-aligned single modality, LASM));
5) 本文提出的标签对齐信息的 MM, 即 LAMM.

3.2 模拟数据集上的实验结果与分析

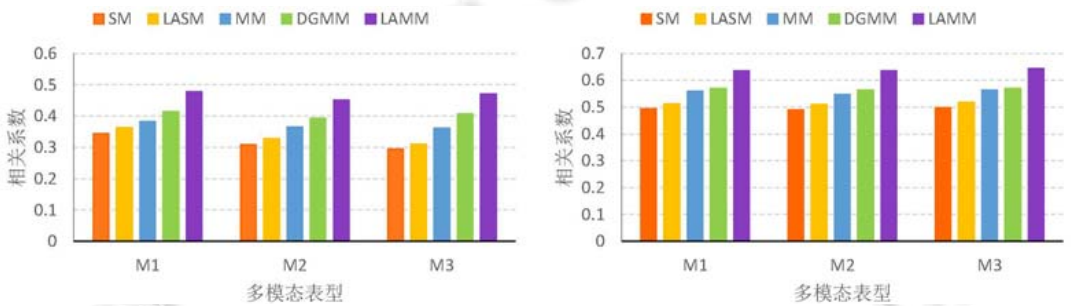
3.2.1 模拟数据集

我们首先利用模拟数据集来评估提出的 LAMM 模型的性能. 此数据集的获取方法可见文献[37], 即: 第 1

步, 生成 p 维向量 $u_t(t=1,\ldots,p)$, 其包括 p'个非零元素; q 维向量 v_k, 其包含 q'个非零元素 $v_{k+1}=v_k+\Delta v$ ($\Delta v\sim N(0,0.1)$), 当 k=1,2,3 时, 得到 3 种不同模态的表型, 记作 $M1$, $M2$ 和 $M3$. u_t 和 v_1 中每一个非零变量从均匀分布中产生, 其范围为[−2,−0.5]∪[0.5,2]. 然后, 随机产生一个样本数为 600 的隐变量 h_1, 其服从正态分布 $N(0,\sigma_{h_1})$, 数据矩阵 Y_1 和 X_1 分别从正态分布 $N(u_th_1,\sigma_eI_q)$和 $N(v_kh,\sigma_eI_q)$中产生. 另外, 随机产生一个样本数为 400 的隐变量 h_2, 其服从正态分布 $N(0,\sigma_{h_2})$, 数据矩阵 Y_2 和 X_2 分别从正态分布 $N(u_th_2,\sigma_eI_q)$和 $N(v_kh,\sigma_eI_q)$中产生. 最后, 我们可以得到样本数为 1 000 和类别数为 2 的数据矩阵 $Y(Y_1\cup Y_2)$和 $X(X_1\cup X_2)$. 在本实验中, 我们假设 p=80, q=100, p'=30, q'=20, $\sigma_{h_1}=\sigma_{h_2}=0.1$.

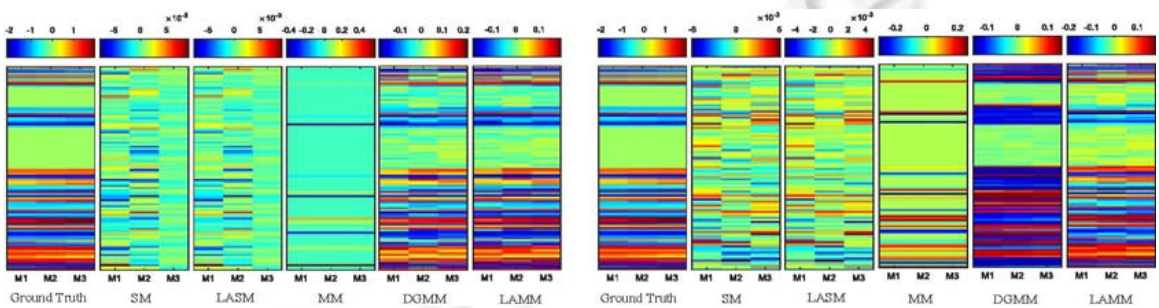
3.2.2 实验结果与分析

在此实验中, 我们将提出的基于标签对齐的方法(包括 LASM 和 LAMM)与基于诊断信息诱导的 DGMM[35]、传统的非标签对齐的(包括 SM 和 MM)[28−33]进行比较. 为了排除偏差, 5 次不同的划分被分别用于 5 次实验中, 并且在上述不同的方法中使用相同的划分. 为了验证不同位点与多模态表型的关联性能, 我们设噪声水平σ_e=0.3, 随机选取 Y 的两个位点 u_{20} 和 u_{40}, 进而产生模拟数据集 1 和模拟数据集 2. 图 2 和图 3 分别给出了 3 种模态回归两个位点 u_{20} 和 u_{40} 的 5 折交叉验证的平均相关系数值以及对应的特征选择结果.



(a) 模拟数据集 1 (b) 模拟数据集 2

图 2 不同方法下, 在模拟数据集上关联的相关性结果



(a) 模拟数据集 1 (b) 模拟数据集 2

图 3 不同方法下, 在模拟数据集上特征选择结果

如图 2 所示, 标签对齐的多模态方法 LAMM 在相关系数的评价指标上一致优于传统的基线方法(SM 和 MM)以及现有研究中最佳性能的 DGMM. 另外, 如图 3 所示: 对于 w 真实值的估计, 相比于其他方法, 提出的 LAMM 方法可以选择出与真实信号更为接近的数据特征, 实现了更高的关联性能, 具备一定优越性.

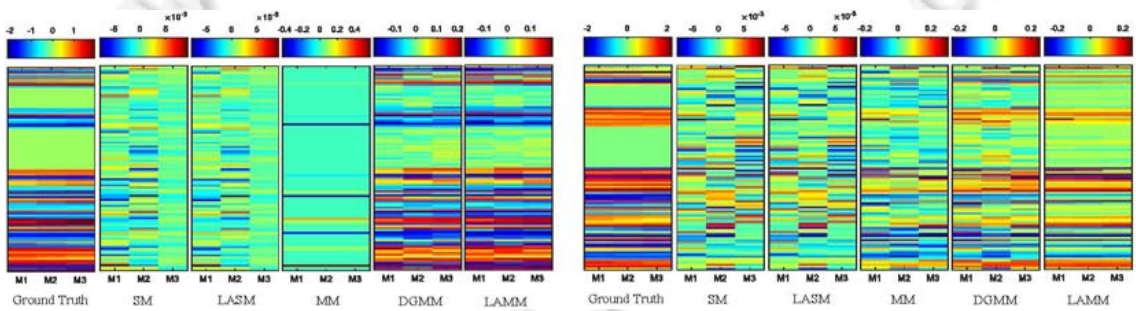
此外, 为了更加全面地评估提出的 LAMM 算法的性能, 我们通过调节噪声水平参数(分别取σ_e 为 0.3 和 0.5), 获得新的模拟数据集 3 和数据集 4. 图 4 和图 5 分别给出了 3 种模态回归位点 u_{20} 的 5 折交叉验证的平均相关系数值以及对应的特征选择结果.

如图 4 所示, 基于标签对齐的回归方法(包括 LASM 和 LAMM)在相关系数的评价指标上一致优于传统的

基线方法(包括 SM 和 MM)以及现有研究中最佳性能的 DGMM. 在图 5 中, 对于估计 w 真实值, SM 和 LASM 方法在不同模态上特征选择结果的分布十分散乱. 另外, 对于噪声较少的模拟数据集 3 来说, DGMM 与 LAMM 具备相互趋同的关联性能. 但是在噪声较强的模拟数据集 4 上, 相比于 DGMM 方法, LAMM 展现出更强的噪声抑制能力和关联性能. 虽然 MM 方法和 DGMM 方法均希望通过嵌入多模态数据之间的互补信息来增强噪声抑制能力, 提高选择更具有关联的判别特征, 但不同模态之间的分布信息仍然没有得到充分利用. 例如: MM 方法仅考虑同一模态在不同任务中的相关性, 忽略了不同模态之间可能存在的内在联系; 而 DGMM 方法通过构建同类样本的相似度矩阵使每一个模态内部的特征在投影后的特征空间中仍能保持类别结构, 通过 $L21$ 范数来保证选择的特征在不同模态上均有较好的判别性能, 但是模态与模态之间的结构信息完全被忽视. 本文提出的 LAMM 方法将不同模态的特征学习当作一个任务, 使用 $L21$ 范数确保只有少数特征能够被联合地从多模态数据中选取. 同时, 使用标号对齐正则化项最大限度地将不同模态之间的类别结构信息嵌入到目标函数中, 增强了噪声抑制的能力, 从而能够诱导出更具有判别性的特征.

(a) 模拟数据集 3　　(b) 模拟数据集 4

图 4　不同方法下在模拟数据集上关联的相关性结果



(a) 模拟数据集 3　　(b) 模拟数据集 4

图 5　不同方法下在模拟数据集上特征选择结果

3.3 ADNI数据集上的实验结果与分析

3.3.1 基因影像数据集

本文采用的 3 种脑影像表型数据(即 VBM, FDG 和 AV45(去除小脑的 90 个脑区))和基因数据(APOE 基因内的 85 个 SNP 位点, 包括与 AD 相关的风险基因位点 APOEe4 SNP rs429358)都来自于 ADNI 数据集. ADNI 数据库的具体相关介绍可以登录 adni.loni.usc.edu 进行查看.

本文中共使用 911 个样本, 包括 160 个 Alzheimer's disease (AD)病人、187 个 late mild cognitive impairment (LMCI)病人、272 个 early mild cognitive impairment (EMCI)病人、82 个 significant memory concern (SMC)病人以及 210 个 Normal Control (NC)正常人, 表 1 给出了这些被试者的统计信息.

表 1 被试者信息统计表

Subjects	NC	SMC	EMCI	LMCI	AD
样本数	210	82	272	187	160
性别(男/女)	109/101	33/49	153/119	108/79	95/65
年龄(均值±方差)	76.13±6.54	72.45±5.67	71.51±7.11	73.86±8.44	75.18±7.88
教育(均值±方差)	16.44±2.62	16.78±2.67	16.07±2.62	16.38±2.81	15.86±2.75
MMSE(均值±方差)	29.00±1.22	29.00±1.22	28.37±1.54	27.71±1.73	24.00±2.62
CDR(均值±方差)	0.01±0.06	0.00±0.00	0.48±0.08	0.49±0.07	0.72±0.27
ADNI-MEM(均值±方差)	1.01±0.57	1.12±0.57	0.59±0.60	0.07±0.67	-0.76±0.61
ADNI-EF(均值±方差)	0.84±0.69	0.73±0.81	0.50±0.74	0.18±0.81	-0.53±0.91
APOE4 阳性(均值±方差)	1.02±0.01	2.15±0.01	3.81±0.02	4.93±0.05	7.82±0.24
Aβt-tau(均值±方差)	2.54±0.08	2.63±0.05	2.45±0.07	2.38±0.06	2.31±0.08

3.3.2 风险基因位点与多模态脑表型关联性能的改进

在此实验中, 我们同样将基于标签对齐的方法(包括 LASM 和 LAMM)、基于诊断信息诱导的 DGMM 和传统的非标签对齐的(包括 SM 和 MM)进行比较. 为了排除偏差, 5 次不同的划分被分别用于 5 次实验中, 并且在上述不同的方法中使用相同的划分. 我们在 VBM, FDG, AV45 这 3 种模态回归 APOE rs429358 数据集上来进行 5 折交叉验证实验, 并将其平均相关系数值统计在表 2 中.

如表 2 所示, 在 3 种模态上, 比较 SM 方法, LASM 的相关系数值为 0.168 6 (VBM), 0.172 1 (FDG), 0.167 2 (AV45), 实现了更优的性能结果. 除此之外, 相比于 DGMM, LAMM 方法在 3 种模态上都获得了更好的相关系数(0.332 1 (VBM), 0.320 1 (FDG), 0.311 4 (AV45)). 这是因为 LASM 使用标号对齐正则化项将类别结构信息嵌入到目标函数中, 诱导出更具有判别性的特征, 则相对于 SM 方法, LASM 取得了更好的性能. 而 MM 方法通过嵌入多模态数据之间的互补信息提高选择更具有关联的判别特征, 则相对于 SM 和 LASM 方法, MM 具有更好的性能. MM 方法和 DGMM 方法均希望通过嵌入多模态数据之间的互补信息来提高选择更具有关联的判别特征, 但不同模态之间的分布信息仍然没有得到充分利用. 例如, MM 方法仅考虑同一模态在不同任务中的相关性, 忽略了不同模态之间可能存在的内在联系. 而 DGMM 方法通过构建同类样本的相似度矩阵使每一个模态内部的特征在投影后的特征空间中仍能保持类别结构, 通过 $L21$ 范数来保证选择的特征在不同模态上均有较好的判别性能, 但是模态与模态之间的结构信息完全被忽视. 本文提出的 LAMM 方法将不同模态的特征学习当作一个任务, 使用 $L21$ 范数确保只有少数特征能够被联合的从多模态数据中选取. 同时, 使用标号对齐正则化项最大限度地将不同模态之间的类别结构信息嵌入到目标函数中, 从而能够诱导出与风险基因位点具有更强关联的判别脑区.

表 2 不同方法的风险基因位点与多模态表型关联的性能比较

方法		相关系数(均值±方差)	
		Tran	Test
SM	VBM	0.0939±0.0096	0.0805±0.0425
	FDG	0.0749±0.0130	0.0596±0.0716
	AV45	0.0536±0.0973	0.0651±0.0908
LASM	VBM	0.2205±0.0164	0.1686±0.0499
	FDG	0.2239±0.0193	0.1721±0.0675
	AV45	0.2564±0.0127	0.1672±0.0422
MM	VBM	0.4407±0.0165	0.2098±0.0497
	FDG	0.4243±0.0197	0.1671±0.0692
	AV45	0.4562±0.0128	0.2276±0.0444
DGMM	VBM	0.3902±0.0222	0.2462±0.0279
	FDG	0.3630±0.0179	0.2301±0.0171
	AV45	0.3990±0.0189	0.2534±0.0269
LAMM	VBM	0.4530±0.0081	**0.3321±0.0082**
	FDG	0.4451±0.0054	**0.3201±0.0094**
	AV45	0.4642±0.0100	**0.3114±0.0062**

此外, 我们也给出了 VBM, FDG, AV45 这 3 种模态回归非风险基因位点数据集上的 5 折交叉验证结果, 平均相关系数值见表 3. 在本文中, 我们选取了一个非 AD 风险基因位点 rs111789331 来验证模型的有效性. 从

表 3 可以看出，每个方法所获得的相关系数都不能反映出基因与脑影像的关联性. 也就是说，关联模型主要面向于包含与疾病相一致的风险基因，通过分离出多模态影像的 QTs 来表示风险基因位点与疾病状态间的内表型特征，从而揭示基因到大脑再到疾病的机理.

表 3　不同方法的非风险基因位点与多模态表型关联的性能比较

方法		相关系数(均值±方差)	
		Tran	Test
SM	VBM	-0.0483±0.0550	0.0113±0.0880
	FDG	0.07071±0.0221	0.02021±0.0543
	AV45	0.0640±0.0435	0.0089±0.0840
LASM	VBM	0.0691±0.0130	0.0502±0.0763
	FDG	0.0271±0.0065	0.0012±0.0508
	AV45	0.0780±0.0021	0.0401±0.0805
MM	VBM	0.3858±0.0148	0.0296±0.0516
	FDG	0.3796±0.0259	0.0977±0.0877
	AV45	0.4021±0.0090	0.0550±0.0498
DGMM	VBM	0.1230±0.0246	0.0636±0.0796
	FDG	0.1619±0.0111	0.0123±0.0410
	AV45	0.1514±0.0517	0.0229±0.0727
LAMM	VBM	0.1618±0.0879	0.0626±0.0826
	FDG	0.2050±0.0635	0.0105±0.0422
	AV45	0.1862±0.0861	0.0401±0.0904

3.3.3　一致的多模态脑区 ROI 的辨别

除了提高风险基因位点与多模态脑表型的关联性能，我们需要找到同时关联风险基因位点和疾病状态的脑区. 在本小节中，通过一致的多模态脑区 ROI 的辨别实验来进一步验证提出算法的优越性. 这里，我们利用不同方法来确定与 APOEe4 rs429358 相关联的所有 ROIs 的权重，其热量图如图 6 所示.

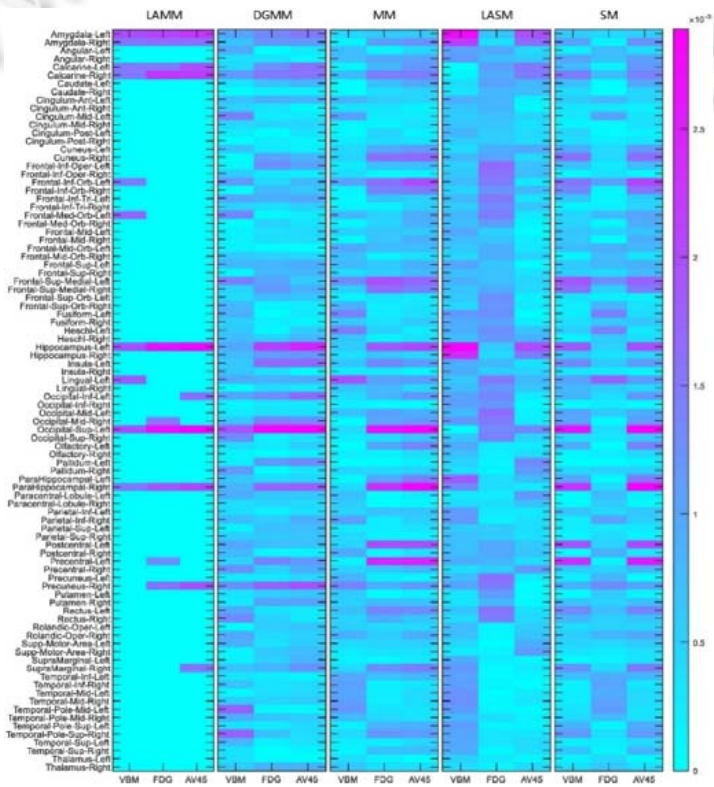


图 6　不同方法与 APOE rs429358 相关联的多模态脑影像 ROIs 特征选择

从该图可知, 提出的LAMM方法能够选择出与APOEe4 rs429358相关联的稀疏的致病ROIs, 这与预期的结果相符. 虽然不同模态脑影像表型和风险基因SNP位点的关联程度有所差异, 但是LAMM可以联合地选取出一致的且与以往的研究相符合的相关脑区ROIs (包括left amygdala, right amygdala, left hippocampus, right parahippocampal gyrus, left calcarine cortex, right calcarine cortex, left superior occipital gyrus 以及 right precuneus)[35,38−40]. 这些结果进一步验证了我们所提出方法的有效性, 并显示了本工作的研究价值和潜力.

3.3.4 最相关脑区ROI的辨别

为了检测鲁棒的脑影像ROI, 表4和图7给出了LAMM所选出的在5折交叉测试中脑影像平均回归系数前10的ROIs以及对应的可视化图. 如预期的那样, 本文提出的LAMM方法已检测到了与风险基因位点相关的前10个ROIs. 值得注意的是, 这些稳定的脑区与之前的研究一致. 即: 在脑影像的ROIs中, left amygdala, right amygdala, left hippocampus, right parahippocampal gyrus, right precuneus, left superior occipital gyrus, right middle occipital gyrus, left calcarine cortex, right calcarine cortex 和 left orbitofrontal cortex (medial)能够作为AD预测和诊断的稳定标志物[35,38−41]. 具体来说, hippocampus 和 parahippocampal gyrus 区域的损伤可以导致记忆的丧失以及定向等功能的障碍[39,40]; amygdala 区域萎缩则与行为的异常有关, 并同时伴随有焦虑和易怒等情绪[38]; right precuneus, left superior occipital gyrus, right middle occipital gyrus, left calcarine cortex, right calcarine cortex 和 left orbitofrontal cortex (medial)均与大脑的代谢改变以及淀粉样蛋白沉积相关[35,41]. 现有的一些研究结果表明: 在临床上, AD的早期阶段会产生大脑的代谢改变以及淀粉样蛋白沉积的现象[42−45], 能够作为AD预测和诊断的标志物.

表4 LAMM方法5折交叉验证得到的前10个重要脑区

ID	ROI	相关文献研究
49	left superior occipital gyrus	[35]
37	left hippocampus	[39]
44	right calcarine cortex	[35]
41	left amygdala	[38]
40	right parahippocampal gyrus	[40]
43	left calcarine cortex	[35]
68	right precuneus	[35]
25	left orbitofrontal cortex (medial)	[41]
42	right amygdala	[38]
52	right middle occipital gyrus	[35]

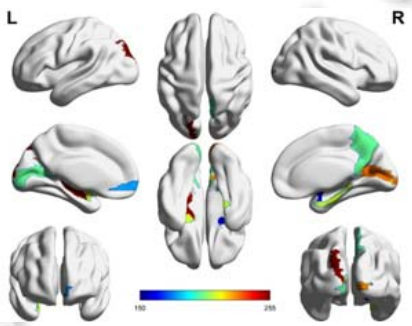


图7 LAMM关联分析中所选择出的前10个ROIs可视化图

4 总 结

针对已有的多模态基因影像关联方法只关注相同样本的多模态信息而忽视了不同模态不同样本间的内在联系, 本文提出了利用多模态脑表型作为桥接风险基因位点和疾病状态的中间特征, 引入标签对齐正则化项, 既充分利用同一模态内部的类别结构信息, 又考虑不同模态之间的信息, 进一步地提高了特征表达能力, 进而能够更为准确地检测出疾病和风险基因位点所关联的脑区特征. 在真实的 ADNI 数据集上, 实验结果表明:

相比已有的多模态基因影像关联方法, 我们的方法在 AD 中发现了风险基因位点与疾病状态之间的鲁棒的一致性表型. 在未来的工作中, 我们将在更多的数据集上评价提出方法的有效性.

References:

[1] Brookmeyer R, Johnson E, Ziegler-Graham K, *et al*. Forecasting theglobal burden of Alzheimer's disease. Alzheimer's and Dementia, 2007, 3(3): 186−191.

[2] Winklern AM, Kochunov P, Blangero J, *et al*. Cortical thicknessor grey matter volume? The importance of selecting the phenotype forimaging genetics studies. NeuroImage, 2010, 53(3): 1135−1146.

[3] Lambert JC ,Ibrahimverbaas CA , Harold D, *et al*. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature Genetics, 2013, 9(4): 1452−1458.

[4] Tian J, Bai J, Bao SL. Preface of the special issue on medical image processing and analysis. Ruan Jian Xue Bao/Journal of Software, 2009, 20(5): 1087−1088 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3616.htm

[5] Zhao XJ, Long ZY, Guo XJ, *et al*. Analysis of magnetic resonance imaging data on the study of Alzheimer's disease. Ruan Jian Xue Bao/Journal of Software, 2009, 20(5): 1123−1138 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3556.htm [doi: 10.3724/SP.J.1001.2009.03556]

[6] Jie B, Zhang DQ. The novel graph kernel for brain networks with application to MCI classification. Chinese Journal of Computers, 2016, 39(8): 1667−1680 (in Chinese with English abstract).

[7] Wang XL, Wang ZQ, Wang ZY, *et al*. Multi-frequency fused graph kernel of brain network for Alzheimer's disease. Chinese Journal of Computers, 2020, 43(1): 64−77 (in Chinese with English abstract).

[8] Zu C. Research on brain image analysis based on sparse structural feature learning and their applications [Ph.D. Thesis]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2017 (in Chinese with English abstract).

[9] Glahn DC, Thompson PM, Blangero J. Neuroimaging endophenotypes: Strategies for finding genes influencing brain structureand function. Human Brain Mapping, 2007, 28(6): 488−501.

[10] Ge T, Schumann G, Feng J. Imaging genetics-towards discoveryneuroscience. Quantitative Biology, 2013, 1(4): 227−245.

[11] Fu Y, Ma Z, Hamilton C, *et al*. Genetic influences on resting-statefunctional networks: A twin study. Human Brain Mapping, 2015, 36(10): 3959−3972.

[12] Hao X, Li C, Yao X, *et al*. Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease. Scientific Reports, 2017, 7: 44272.

[13] Hao X, Li C, Yan J, *et al*. Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis. Bioinformatics, 2017, 33(14): i341−i349.

[14] Song A, Yan J, Kim S, *et al*. Network-based analysis of genetic variants associated with hippocampal volume in Alzheimer's disease: Astudy of ADNI cohorts. BioData Mining, 2016, 9(1): 1−8.

[15] Yao X, Yan J, Liu K, *et al*. Tissue-specific network-based genomewide study of amygdala imaging phenotypes to identify functionalinteraction modules. Bioinformatics, 2017, 33(20): 3250−3257.

[16] Consortium B, Anttila V, Bulik-Sullivan B, *et al*. Analysis of sharedheritability in common disorders of the brain. Science, 2018, 360(6395): eaap8757.

[17] Shao W, Han Z, Cheng J, *et al*. Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis. IEEE Trans. on Medical Imaging, 2019, 39(1): 99−110.

[18] Wang M, Shao W, Hao X, *et al*. Identify consistent cross-modality imaging genetic patterns via discriminant sparse canonical correlation analysis. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2021, 18(4): 1549−1561.

[19] Shao W, Xiang S, Zhang Z, *et al*. Hypergraph based sparse canonical correlation analysis for the diagnosis of Alzheimer's disease from multi-dimensional genomic data. Methods, 2021, 189: 86−94.

[20] Hao XK. Research on machine-learning-based imaging genetics analysis and their applications [Ph.D. Thesis]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2017 (in Chinese with English abstract).

[21] Consortium TGP. A global reference for human genetic variation, the 1000 genomes project consortium. Nature, 2015, 526: 68−74.

[22] Stein JL, Hua X, Lee S, *et al*. Voxelwise genome-wide association study (vGWAS). NeuroImage, 2010, 53(3): 1160−1174.

[23] Hibar DP, Stein JL, Kohannim O, *et al*. Voxelwise gene-wide association study (vGeneWAS): Multivariate gene-based association testing in 731 elderly subjects. NeuroImage, 2011, 56(4): 1875−1891.

[24] Brun CC, LeporeN, Pennec X, *et al*. Mapping the regional influence of genetics on brain structure variability—A tensor-based morphometry study. NeuroImage, 2009, 48(1): 37−49.

[25] Filippini N, Rao A, Wetten S, *et al*. Anatomically-distinct genetic associations of APOE epsilon4 allele load with regional cortical atrophy in Alzheimer's disease. NeuroImage, 2009, 44(3): 724−728.

[26] Baranzini SE, Wang J, Gibson RA, *et al*. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Human Molecular Genetics, 2009, 18(4): 767−778.

[27] Potkin SG, Turner JA, Guffanti G, *et al*. Genome-Wide strategies for discovering genetic influences on cognition and cognitive disorders: Methodological considerations. Cognitive Neuropsychiatry. 2009, 14(4): 391−418.

[28] Vounou M, Nichols TE, Montana G. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. NeuroImage, 2010, 53(3): 1147−1159.

[29] Vounou M, Janousova E, Wolz R, *et al*. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. NeuroImage, 2012, 60(1): 700−716.

[30] Batmanghelich NK, Dalca AV, Sabuncu MR, *et al*. Joint modeling of imaging and genetics. Inf Process Med Imaging, 2013, 23(23): 766−777.

[31] Argyriou A, Evgeniou T, Pontil M. Convex multi-task feature learning. Machine Learning, 2008, 73(3): 243−272.

[32] Obozinski G, Taskar B, Jordan MI. Joint covariate selection and joint subspace selection for multiple classification problems. Statistics and Computing, 2010, 20(2): 231−252.

[33] Zhang D, Shen D. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage, 2012, 59(2): 895−907.

[34] Shen L, Thompson PM. Brain imaging genomics: Integrated analysis and machine learning. Proc. of the IEEE, 2020, 108(1): 125−162.

[35] Hao X, Yao X, Yan J, *et al*. Identifying multimodal intermediate phenotypes between genetic risk factors and disease status in Alzheimer's disease. Neuroinformatics, 2016, 14(4): 1−14.

[36] Chen X, Pan WK, Kwok JT, *et al*. Accelerated gradient method for multi-task sparse learning problem. In: Proc. of the 9th IEEE Int'l Conf. on Data Mining. Miami, 2009. 746−751.

[37] Fang J, Lin D, Schulz SC, *et al*. Joint sparse canonical correlation analysis for detecting differential imaging genetics modules. Bioinformatics, 2016, 32(22): 3480−3488.

[38] Horinek D, Varjassyova A, Hort J. Magnetic resonance analysisof amygdalar volume in Alzheimer's disease. Current Opinion in Psychiatry, 2007, 20(3): 273−277.

[39] Laakso MP ,Frisoni GB , Knnen M, *et al*. Hippocampus and entorhinal cortex in frontotemporal dementia and Alzheimer's disease: A morphometric MRI study. Biological Psychiatry, 2000, 47(12): 1056−1063.

[40] Shen L, Kim S, Risacher SL, *et al*. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. NeuroImage, 2010, 53(3): 1051−1063.

[41] Du L, Huang H, Yan J, *et al*. Structured sparse canonical correlationanalysis for brain imaging genetics: An improved graphnet method. Bioinformatics, 2016, 32(10): 1544−1551.

[42] Liu Y, Yu JT, Wang HF, *et al*. APOE genotype and neuroimaging markers of Alzheimer's disease: Systematic review and meta-analysis. Journal of Neurology Neurosurgery & Psychiatry, 2015, 86(2): 127−134.

[43] Reiman EM , Caselli RJ , Yun LS , *et al*. Preclinical evidence of Alzheimer's disease in persons homozygous for the epsilon4 allele for apolipoprotein E. New England Journal of Medicine, 1996, 334(12): 752−758.

[44] Camus V, Payoux P, Barre L, *et al*. Using PET with 18F-AV-45 (florbetapir) to quantify brain amyloid load in a clinical environment. European Journal of Nuclear Medicine and Molecular Imaging, 2012, 39(4): 621−631.

[45] Wishart HA, Saykin AJ, Mcallister TW, *et al*. Regional brain atrophy in cognitively intact adults with a single APOE epsilon4 allele. Neurology, 2006, 67(7): 1221−1224.

附中文参考文献:

[4] 田捷, 白净, 包尚联. 医学影像处理与分析专刊前言. 软件学报, 2009, 20(5): 1087−1088. http://www.jos.org.cn/1000-9825/3616.htm

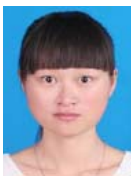
[5] 赵小杰, 龙志颖, 郭小娟, 等. 阿尔茨海默氏症研究中的磁共振成像数据分析. 软件学报, 2009, 20(5): 1123−1138. http://www.jos.org.cn/1000-9825/3556.htm [doi: 10.3724/SP.J.1001.2009.03556]

[6] 接标, 张道强. 面向脑网络的新型图核及其在 MCI 分类上的应用. 计算机学报, 2016, 39(8): 1667−1680.

[7] 汪新蕾, 王之琼, 王中阳, 等. 面向阿尔茨海默病的脑网络多频段融合图核. 计算机学报, 2020, 43(1): 64−77.

[8] 祖辰. 基于稀疏结构特征学习的脑图像分析及其应用研究 [博士学位论文]. 南京: 南京航空航天大学, 2017.

[20] 郝小可. 基于机器学习的影像遗传学分析及其应用研究 [博士学位论文]. 南京: 南京航空航天大学, 2017.



汪美玲(1988－), 女, 博士, CCF 专业会员, 主要研究领域为影像遗传学, 机器学习.

邵伟(1986－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为生物信息学, 机器学习.

张道强(1978－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为医学影像分析, 数据挖掘, 机器学习.