

基于机器学习的脑影像基因组学分析方法综述

汪美玲¹, 刘青山¹, 张道强^{2,3}

(1. 南京邮电大学计算机学院, 南京 210023; 2. 南京航空航天大学人工智能学院, 南京 211106; 3. 南京航空航天大学脑机智能技术教育部重点实验室, 南京 211106)

摘要: 脑影像基因组学是一个新兴的数据科学领域。在该领域中通过对脑影像数据与基因组数据(通常还结合其他生物标志物、临床数据及环境数据)进行综合分析,可以深入探究大脑的表型、遗传及分子特征,以及这些特征对正常和异常脑功能及行为的影响。鉴于机器学习在生物医学领域的作用日益重要,且脑影像基因组学相关文献迅速增长,本文对脑影像基因组学中机器学习方法进行了最新且全面的综述。本文首先回顾了脑影像基因组学的相关背景和基础工作;然后展示了基于多变量机器学习的脑影像基因组学关联研究的主要思想和建模,并提出了联合关联分析和结果预测的方法;最后对今后的工作进行了展望。

关键词: 机器学习; 脑影像基因组学; 关联分析; 智能诊断; 脑疾病

中图分类号: TP183

文献标志码: A

A Review of Machine Learning for Brain Imaging Genomic Analysis

WANG Meiling¹, LIU Qingshan¹, ZHANG Daoqiang^{2,3}

(1. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; 2. College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; 3. Key Laboratory of Brain-Machine Intelligence Technology of Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Brain imaging genomics is a burgeoning domain within data science, where an integrated analytical approach is applied to brain imaging and genomics data, frequently in conjunction with other biomarker, clinical, and environmental datasets. This strategy is employed to glean fresh insights into the phenotypic, genetic, and molecular features of the brain, along with their effects on both typical and atypical brain function and behavior. In light of the escalating significance of machine learning in biomedicine and the swiftly expanding corpus of literature in brain imaging genomics, this paper presents a current and exhaustive review of machine learning methodologies tailored for brain imaging genomics. Firstly, the related background and fundamental work in imaging genomics are reviewed. Then, we summarize the main idea and modelling in genetic-imaging association studies based on multivariate machine learning and present methods for joint association analysis and outcome prediction. Finally, this paper discusses some prospects for future work.

Key words: machine learning; brain imaging-genomics; association analysis; intelligent diagnosis; brain disease

基金项目: 国家自然科学基金(62136004, 62106104, 62276130); 国家重点研发计划(2024YFC3308402); 南京邮电大学引进人才自然科学研究启动基金(NY223170); 南京邮电大学校级自然科学基金(NY224115); 中央高校基本科研业务费项目(NJ2024029)。

收稿日期: 2025-06-21; **修订日期:** 2025-07-23

引言

近年来,随着认知神经科学的发展,神经影像学为人类大脑工作机制的研究注入了新的活力。与此同时,随着非侵入性脑成像技术的进步,研究人员期望能够更深入地了解大脑的成像特征和分子机制,以及它们对正常和异常脑功能及行为的影响。常用的脑成像技术包括结构磁共振成像(Structural magnetic resonance imaging, sMRI)、功能磁共振成像(Functional magnetic resonance imaging, fMRI)、弥散张量成像(Diffusion tensor imaging, DTI)和正电子发射断层成像(Positron emission tomography imaging, PET)。此外,随着遗传技术的发展,研究人员能够在更精细的分子层面,如单核苷酸多态性(Single nucleotide polymorphisms, SNPs),识别与神经和精神疾病相关的遗传标记。

由于获取多模态脑成像数据和高通量基因组数据方面的技术不断进步,脑影像基因组学(Brain imaging genomics)正迅速崛起为一个新兴的研究领域。Thompson等^[1]提出了影像基因组学(Imaging genomics)或影像遗传学(Imaging genetics)的概念,通过开展整合性研究,分析遗传变异(如SNPs)、表观遗传变异和拷贝数变异(Copy number variations, CNVs)、由各类组学数据捕获的分子特征,以及脑成像定量特征(Quantitative traits, QTs),同时结合其他生物标志物、临床和环境数据进行分析^[1-2]。脑影像基因组学领域的诸多进展,得益于多项大规模标志性研究,如阿尔茨海默病神经影像学计划(Alzheimer's disease neuroimaging initiative, ADNI)^[3]、通过元分析强化神经影像遗传学(Enhancing neuro imaging genetics through meta analysis, ENIGMA)联盟^[4]以及英国生物银行(UK Biobank)^[5]。这些研究为全球科研界提供了大规模脑影像基因组数据,推动了该领域方法学创新与生物医学应用丰富的文献产出,其中包括多篇从不同视角总结相关进展的综述文章。例如,ADNI是阿尔茨海默病(Alzheimer's disease, AD)生物标志物研究的里程碑式项目。ADNI队列构建了极为丰富的多模态数据资源库,涵盖全基因组分型、全基因组测序、血液转录组、血液表观基因组、血浆/血清/脑脊液蛋白质组、血浆/血清代谢组、多模态磁共振成像(Magnetic resonance imaging, MRI)与正电子发射断层扫描PET等神经影像数据,以及认知、行为和临床数据。基于其开放科学特性,ADNI数据已被全球科研界广泛用于脑影像基因组学研究,催生了数百篇相关论文。ADNI遗传学核心组^[6-7]及整个ADNI团队^[3,8]定期对这些进展进行综述。ENIGMA是另一项对脑影像基因组学领域产生重大影响的重大倡议。ENIGMA联盟作为全球性科研协作网络,以揭示疾病与遗传因素对大脑的影响为共同目标。ENIGMA联盟的进展已在多篇综述文章中被定期总结^[4,9-11]。Thompson等^[4]最新更新的ENIGMA联盟进展显示,该联盟已汇聚来自43个国家的1400余名科学家,共同开展基于影像、基因组学及其他脑指标的人类大脑研究。UK Biobank作为一项前瞻性流行病学队列研究,覆盖超50万个体,是另一项提供海量脑影像基因组数据的重要研究^[5],其已公开了约50万例样本的全基因组数据^[12],以及约1.5万例样本的六模态全脑影像数据^[13]。Elliott等^[14]近期完成了大规模脑影像定量特征全基因组关联分析,在8428例样本的发现集(针对3144个影像QTs检测超1100万个SNPs)及930例和3456例样本的两套独立验证集中开展研究。其作为大规模脑影像基因组学的前沿成果,为揭示大脑遗传结构提供了重要的研究视角。

脑影像基因组学主要借助脑影像技术,将脑结构及功能视作表型指标来评估遗传因素对个体产生的作用,同时深入探究基因影响大脑神经结构与功能的方式,以及由此导致的神经病理学改变情况。通过研究遗传因素和脑结构、功能之间的内在联系,在“基因与大脑”之间构建起一座直观的桥梁,这对于深入揭示神经精神疾病的发病机理具有重要意义^[15-17]。影像基因组学还能识别脑疾病的生物学指标或内表型,为疾病的预测和诊断提供更准确的方法。具体而言,大多数研究者将SNP作为关联分析的基因型数据。在内表型数据的获取方面,研究者多采用临床脑影像数据(如磁共振成像)进行分析。例如,sMRI作为一种测量大脑结构组织的成像技术,可量化形态学(如灰质体积)的异常;fMRI扫描则被证实能有效揭示大脑的功能连接模式。基于不同模态的脑影像技术,目前影像基因组学主要聚焦于基

因 SNP 与大脑结构、功能及连接性之间的关联分析^[18-21]。

早期的影像基因组学方法主要采用单变量配对统计分析方法,通过多重检验来寻找 SNP 或基因与复杂疾病或可测量定量性状 QTs 之间的关联。全基因组关联研究 (Genome-wide association study, GWAS) 利用全基因组高通量测序技术对研究对象的基因组序列变异进行分类,最终通过生物统计学方法和生物信息学方法筛选出具有显著意义的 SNP^[22]。自 2005 年《Science》杂志发表首篇关于年龄相关性黄斑变性的 GWAS 研究论文以来^[23],该方法已被广泛应用于精神疾病的分析中^[24]。GWAS 在影像遗传学研究中发挥了重要作用,但也存在一些问题,如严格的多重校正导致许多小效应变异无法通过校正水平。此外,GWAS 仅能获得遗传变异与性状之间的单一关联度,无法很好地解释大脑复杂的分子机制。

近年来,随着机器学习理论的快速发展,研究者开始尝试利用这些数据分析工具解决多个领域的问题。在影像遗传学的关联分析中,除了单变量统计分析外,多变量机器学习模型应用最为广泛,并已成功识别出与疾病敏感的影像和遗传生物标志物。已有学者发表了影像遗传学相关方法的综述论文。例如,Liu 等^[25]回顾了用于分析和整合影像与遗传数据的多元方法。郝小可和 Wang 等^[26-27]回顾了脑影像基因组学的回归与相关性方法,以及用于挖掘高级影像基因组关联的基于集合的方法。Mufford 等^[28]回顾了精神病学领域脑影像基因组学的方法。简而言之,这些综述从不同角度涵盖了脑影像基因组学的主题,其中部分综述侧重于回顾特定研究(如 ADNI^[3,6-8]或 ENIGMA^[4,9-11])的数据、方法、分析和/或结果,部分综述考察了特定学科(如文献[17]中的精神病学)或疾病(如文献[28]中的精神分裂症)背景下的研究活动和进展,还有一些综述则提供了关于多变量分析^[26]和机器学习^[27]的方法导向性回顾。鉴于统计学和机器学习在生物医学研究中发挥着越来越重要的作用,新方法不断涌现^[29],因此本文全面介绍了用于解决脑影像基因组学问题的最新机器学习方法,并就各种生物医学应用的方法选择进行实际讨论。图 1 展示了脑影像基因组学所涵盖主题的示意图。基于机器学习的影像基因组学分析方法的主要目标之一是实现关联分析研究。本文将这些基于机器学习的影像基因组学方法分为两类:第一类主要利用回归模型来识别复杂的多 SNP 和/或多 QT 关联;第二类为脑影像基因组学研究开发的重要方法,如稀疏典型相关分析 (Sparse canonical correlation analysis, SCCA)^[30-32]和平行独立成分分析 (Parallel-independent component analysis, pICA)^[33-34]。总体而言,本文聚焦于以下 3 类学习问题进行综述:首先,阐述多变量影像遗传学关联分析的问题,并总结基于多变量机器学习的遗传-影像关联研究的主要思想和建模方法;其次,回顾结合影像和基因组数据来预测感兴趣结果的方法,以及联合关联分析和结果预测的方法;最后,展望了脑影像基因组学研究领域尚未解决的问题及未来的研究方向。

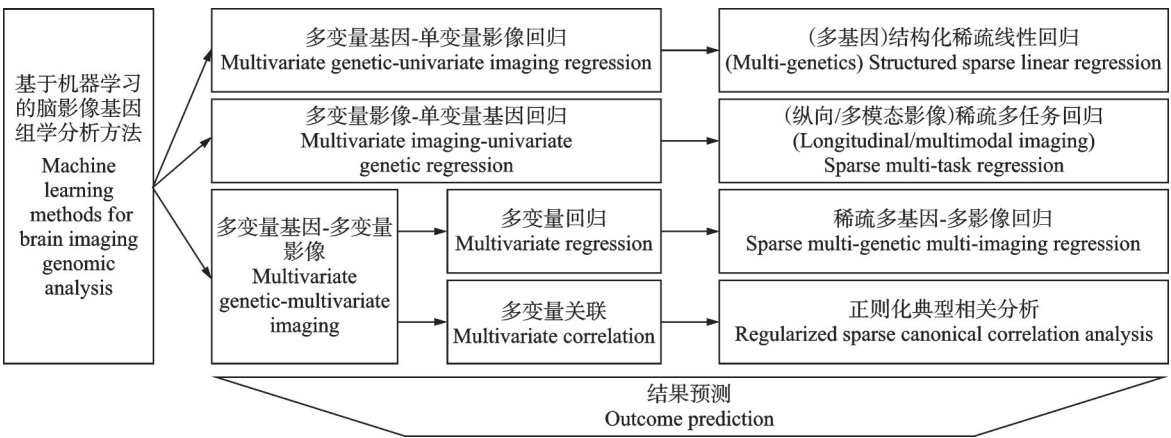


图 1 基于机器学习的脑影像基因组学分析方法

Fig.1 Machine learning for brain imaging genomic analysis

1 多变量回归

本节对采用回归模型来识别复杂的多 SNP 和/或多 QT 关联的机器学习研究进行综述。设 $X \in \mathbb{R}^{n \times p}$ 为包含 n 个受试者、 p 个变量的遗传数据, $Y \in \mathbb{R}^{n \times q}$ 为包含 n 个受试者、 q 个变量的影像数据。本文假定 X 和 Y 的每一列均已标准化,均值为零,方差为单位方差。在脑影像基因组学研究中大多数回归模型均可用以下通用正则化损失函数框架来描述,即

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{X}, \mathbf{Y}, \mathbf{W}) + \sum_{i=1}^m \lambda_i R_i(\mathbf{W}) \quad (1)$$

式中: $\mathbf{W} \in \mathbb{R}^{p \times q}$ 是 Y 对 X 进行回归的权重矩阵; λ_i 是用于平衡损失函数 $\mathcal{L}(\mathbf{X}, \mathbf{Y}, \mathbf{W})$ 和正则化项 $R_i(\mathbf{W})$ 的参数。这些模型中通常会纳入一项能诱导稀疏性的正则化项。其动机有两方面:首先,合理假设最终影像基因组关联中仅有少量标记物是相关的,稀疏项有助于识别这些相关标记物;其次,稀疏性约束可以降低模型复杂度,进而降低过拟合风险。

1.1 多变量基因-单变量影像回归

采用稀疏正则化回归模型来实现多变量基因-单变量影像回归的主要动机有两点。首先,假设仅有少数标记物与影像基因组学相关联,稀疏项有助于识别这些相关标记物。其次,稀疏约束可降低模型复杂度及过拟合风险。在文献[35-36]中,基于 L_1 范数惩罚约束的回归模型已成功应用于多变量基因数据分析,旨在识别与特定脑区高度关联的稀疏 SNP 位点。这些模型为检测和识别高维基因 SNP 的小样本回归问题提供了通用技术框架。然而,基于 L_1 范数的约束并未充分考虑特征变量间的结构关系,因此理论上无法获得最优回归结果。考虑到 SNP 特征间的空间结构关系, Silver 和 Yuan 等^[37-39]提出了组稀疏模型或融合稀疏模型,用于选择同一组或相邻特征变量中的 SNP 位点,基于组稀疏性或融合稀疏的模型如下

$$\min_{\mathbf{w}} \left\| \mathbf{y} - \mathbf{X}\mathbf{w} \right\|_2^2 + \lambda \sum_{i=1}^g \sqrt{\sum_{j \in G(i)} \mathbf{w}_j^2} \quad (2)$$

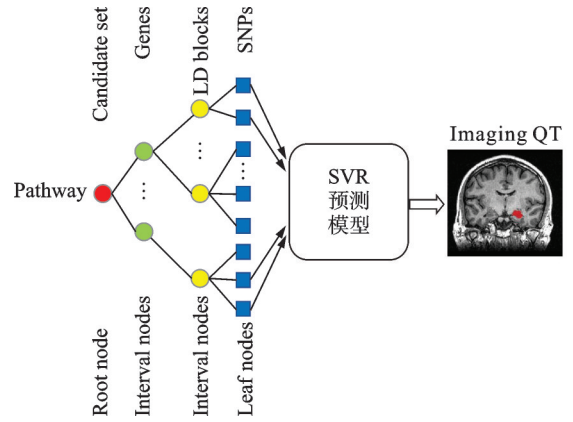
$$\min_{\mathbf{w}} \left\| \mathbf{y} - \mathbf{X}\mathbf{w} \right\|_2^2 + \lambda \sum_{i < j} |\mathbf{w}_i - \mathbf{w}_j| \quad (3)$$

式(2~3)分别用于从基因数据中识别一组 SNP 位点 X , 并预测单一成像表型 y 。在式(2)中,组稀疏项中的 $\mathbf{w}_j$ 表示属于该组的所有 SNP 位点特征,其目标在于控制所选位点以体现聚类特性。例如,基因位点间存在连锁不平衡(Linkage disequilibrium, LD)效应^[40],即不同基因上连锁的 SNP 会非随机地出现在同一 LD 区块中。这为基于组稀疏性的特征选择模型提供了领域知识,使得同一 LD 组中的 SNP 可被同时检测出来。在式(3)中,融合 Lasso 项可控制相邻位置特征的 $\mathbf{w}_i$ 和 $\mathbf{w}_j$ 权重贡献,使其尽可能相似,即通过融合 Lasso 项选择的特征变量具有空间连续性。

此外, SNP 位点之间不仅存在平面空间关系,在实际基因结构中还存在层次关系。例如,在特定信号通路中,特定基因位点的相互作用会影响蛋白质合成与功能转化,且同一基因下的某些 SNP 位点也存在特定相关性(如连锁不平衡 LD)。因此,充分利用这种层次结构的先验知识进行影像遗传学分析,往往能降低回归分析误差,并学习到更具解释性的特征模式^[41-43],如图 2 所示。该模型采用树引导稀疏学习(Tree-guided sparse learning, TGSL)方法识别基因型与表型间的关联。构建树结构时,以 SNP 位点作为叶节点, LD 区块和基因区块作为中间节点,信号通路中的所有基因作为最终根节点。该结构树具有 d 层,每层均包含 n_i 节点。设第 i 层是 $\{G_1^i, G_2^i, \dots, G_j^i, \dots, G_{n_i}^i\}$, 则树引导稀疏回归模型如下

$$\min_{\mathbf{w}} \left\| \mathbf{y} - \mathbf{X}\mathbf{w} \right\|_2^2 + \lambda \sum_{i=1}^d \sum_{j=1}^{n_i} \alpha_j^i \left\| \mathbf{w}_{G_j^i} \right\|_2 \quad (4)$$

式中: α_j^i 代表根据先验知识预先设定的任意节点 G_j^i 的权重; $w_{G_j^i}$ 代表所学习到的树结构中任意节点 G_j^i 的权重。该研究旨在确定一组用于预测单一成像表型 y 的 SNP 位点。值得注意的是, 当一个节点的权重为零时, 其所有子节点的权重也均为零, 即该子树的所有特征均与回归任务无关, 因此不会被选中。与传统 Lasso 方法相比, 通过模型优化得到的 SNP 在预测大脑灰质体积时误差更小, 且这些与 MRI 脑区相关的 SNP 位点具有层次聚类特性。该研究可以识别块级稀疏 SNP 模式, 从而为生物学解释提供更好的指导。

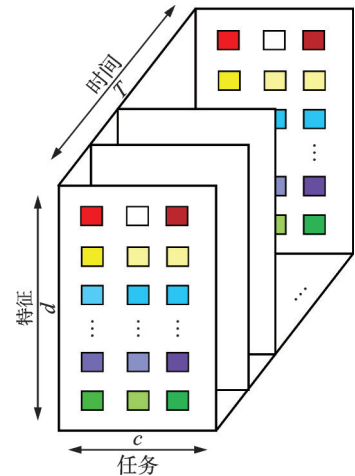
图2 树引导的稀疏回归模型^[41-43]Fig.2 Tree-guided sparse regression model^[41-43]

1.2 多变量影像-单变量基因回归

在基于机器学习的影像遗传学研究中, 大部分工作都聚焦于发现和检测与影像表型相关的多变量 SNP 位点。然而, 鲜有研究探讨当表型测量变量发生变化时, SNP 值如何随之改变, 即利用多变量影像数据对单变量基因特征进行回归分析。例如, Shen 等^[44]提出了一种基于预测变量间组结构信息的任务相关时间序列多变量稀疏回归模型, 表示如下

$$\min_{\mathbf{W}} \sum_{t=1}^T \|\mathbf{Y} - \mathbf{X}_t^T \mathbf{W}_t\|_F^2 + \lambda_1 \sum_{k=1}^d \sqrt{\sum_{t=1}^T \|\mathbf{w}_t^k\|_2^2} + \lambda_2 \text{Tr}(\mathbf{W} \mathbf{W}^T)^{\frac{1}{2}} \quad (5)$$

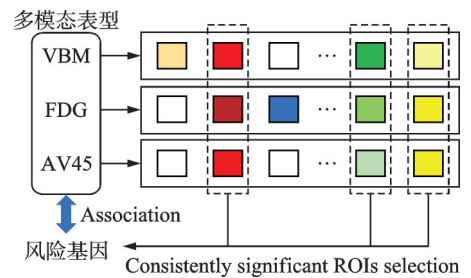
式中: 脑影像数据为 $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T\} \in \mathbb{R}^{d \times n \times T}$, 其中 d 表示影像特征维数, n 表示样本数, T 表示时间; SNP 数据为 $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\} \in \mathbb{R}^{n \times c}$, 其中 c 表示任务数; $\mathbf{w}_t^k$ 表示在时间 t 时系数矩阵 $\mathbf{W}_t$ 的第 k 行; $\text{Tr}(\cdot)$ 为矩阵的迹; 张量系数矩阵为 $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_T\} \in \mathbb{R}^{d \times c \times T}$, 揭示了脑成像定量特征 QTs 的时间序列。当 $c=1$ 时, 该算法用于分析时间序列影像 MRI 与风险基因之间的关联, 如图 3 所示。通过多回归任务和多个时间点上对特征权重进行联合约束, 可以筛选出与任务相关的纵向成像表型标志物。该模型从表型到基因型分析提供了一个新视角, 用于研究单个基因对大脑结构和功能变化的影响。

图3 任务相关纵向稀疏回归模型^[44]Fig.3 Task-correlated longitudinal sparse regression model^[44]

此外, 在单变量基因-多变量影像关联分析中, 大多数研究聚焦于单一模态影像表型 QT。为研究遗传因素与多模态脑影像定量特征之间的关联, Hao 等^[45-46]通过引入组稀疏正则化项, 构建了诊断信息引导的多模态 (Diagnosis-guided multimodal, DGMM) 回归模型 (图 4), 实现了多模态影像定量特征 $\mathbf{Y}$ 与候选风险基因位点 x 之间的关联分析, 表达式为

$$\min_{\mathbf{W}} \sum_{m=1}^M \|\mathbf{x} - \mathbf{Y}_m \mathbf{w}_m\|_2^2 + \lambda_1 \sum_{i,j} \sqrt{\sum_j \mathbf{w}_{ij}^2} + \lambda_2 R(\mathbf{w}) \quad (6)$$

式中: $\mathbf{W} = [\mathbf{w}_{\text{VBM}}, \mathbf{w}_{\text{FDG}}, \mathbf{w}_{\text{AV45}}]$ 为多模态影像定量特征

图4 多模态关联模型^[45-46]Fig.4 Multi-modality association model^[45-46]

(VBM、FDG、AV45)与候选风险基因 APOE $\epsilon 4$ SNP 的关联权重矩阵,其中 VBM 是通过结构磁共振成像数据进行预处理得到的基于体素的形态学测量值,FDG 为氟脱氧葡萄糖正电子发射断层扫描,AV45 为 PET 扫描淀粉样蛋白成像;第 2 项为组稀疏正则化项; $R(\mathbf{w})$ 为拉普拉斯正则化项。该方法通过广义线性回归函数,实现与风险基因相关的多模态成像生物标志物的特征选择。与单模态关联方法相比,该多模态关联方法能够识别出稳健一致的脑区,且具有较强的抗噪声能力。因此,该方法可应用于其他风险基因与多模态成像定量特征之间的关联分析。该实证研究基于阿尔茨海默病神经影像学倡议 ADNI 样本开展,其中响应变量为 APOE $\epsilon 4$ SNP,预测变量包括 3 种模态的感兴趣区域 ROI 测量值:(1) 结构磁共振成像的 VBM 测量值;(2) 正电子发射断层扫描 PET 的 FDG 测量值;(3) PET 的 AV45 测量值。例如,Wang 等^[47]提出了一种诊断对齐的多模态 (Diagnosis-aligned multimodal, DAMM) 策略,用于候选风险基因 APOE $\epsilon 4$ SNP 的 x 对多模态成像定量特征 Y (结构磁共振成像的 VBM 测量值和基于 fMRI 的超图聚类系数测量值) 的回归分析,具体如下

$$\min_{\mathbf{w}} \sum_m \left\| \mathbf{x} - \mathbf{Y}_m \mathbf{w}_m \right\|_2^2 + \lambda_1 R_1(\mathbf{w}) + \lambda_2 R_2(\mathbf{w}) \quad (7)$$

式中: R_1 是与式(5)相同的组稀疏正则化项,目的是在多模态脑成像定量特征中联合筛选出与风险 SNP 位点相关的少量特征; R_2 是图拉普拉斯正则化项,用于充分利用不同模态间的信息。该模型将结构体素信息与网络连接信息作为连接风险基因位点与疾病状态的中间特征,以探寻风险 SNP 与脑网络之间针对特定疾病的关联。值得注意的是,这项研究是探索连通性特征与遗传变异关系的初步尝试。该实证研究基于阿尔茨海默病神经影像学倡议 ADNI 样本开展,其中响应变量为 APOE $\epsilon 4$ SNP,预测变量包括两种模态的感兴趣区域 ROI 测量值:(1) 结构磁共振成像的 VBM 测量值;(2) 基于 fMRI 的超图聚类系数测量值。

1.3 多变量基因-多变量影像回归

上述针对 SNP 位点的多变量分析仅针对影像单变量特征的回归,无法充分利用影像多变量特征之间的相关性。近年来,已引入能够处理多变量基因输入和多变量影像输出的高维基因与影像数据分析方法。部分研究采用稀疏回归方法,从原始高维数据中发现与影像 QT 显著相关的低维基因数据集^[48]。例如,Vounou 等^[49]提出了稀疏降秩回归 (Sparse reduced-rank regression, SRRR) 模型,表达式为

$$\min_{A,B} \left\| \mathbf{Y} - \mathbf{XBA}^T \right\|_F^2 + \lambda_1 \left\| \mathbf{A} \right\|_1 + \lambda_2 \left\| \mathbf{B} \right\|_1 \quad (8)$$

式中 $\mathbf{W}$ 表示矩阵 $\mathbf{B}$ (大小为 $p \times r$) 与矩阵 $\mathbf{A}$ (大小为 $q \times r$) 的乘积,即 $\mathbf{W} = \mathbf{BA}^T$ 。该模型的主要目标是求矩阵 $\mathbf{W}$ 秩的最小值。对矩阵 $\mathbf{A}$ 和 $\mathbf{B}$ 施加 L_1 范数约束以实现特征稀疏选择。图 5 展示了稀疏降秩回归模型的可视化框架,其中 ROI 表示感兴趣区域。

进一步,Vounou 等^[49]对其先前的研究进行了改进。具体而言,首先在结构性脑成像定量特征中采用判别分析方法,筛选出具有判别能力的多变量生物标志物作为表型,随后利用现有的稀疏降秩回归 SRRR 模型开展全基因组关联分析,最终取得了良好效果^[50]。随后,Silver 等^[51]提出了一种通路稀疏降秩回归 (Pathways SRRR, P-SRRR) 方法,该方法将通路组稀疏与自适应权重 (Pathways group lasso with adaptive weights, P-GLAW) 理念融入 SRRR 方法中,即

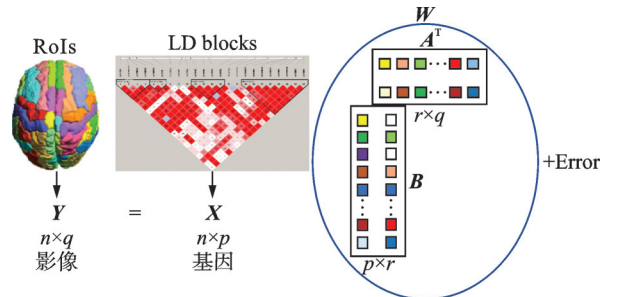


图 5 稀疏降秩回归模型^[49]

Fig.5 Sparse reduced rank regression model^[49]

$$\min_{A, B} \|Y - XBA^T\|_F^2 + \lambda \sum_{g \in G} d_g \|B_g\| \quad (9)$$

式中: G 被定义为矩阵 B 的分组结构; d_g 表示权重; B_g 表示基因位点 SNP 系数对应的向量。该模型从 X 中识别出一组 SNP, 从而能够预测一组与脑疾病相关的影像定量特征 Y 。该方法利用通路知识对 SNP 进行分组, 以便在通路层面进行特征选择。

此外, 部分研究还考虑了多变量多任务回归模型。例如, Wang 等^[52]提出了一种组稀疏多任务回归与特征选择 (Group-sparse multi-task regression and feature selection, G-SMuRFS) 策略 (图 6), 该策略是一种结构化稀疏模型, 表达式为

$$\min_W \|Y - XW\|_F^2 + \lambda_1 \sum_{k=1}^K \sqrt{\sum_{i \in \pi_k} \sum_j w_{ij}^2} + \lambda_2 \sum_i \sqrt{\sum_j w_{ij}^2} \quad (10)$$

式中: 第 2 项考虑了 SNP 位点间的连锁不平衡 LD 结构关系, 嵌入了 SNP 分组关系的先验信息, 从而可同时检测同一连锁不平衡组中的 SNP; 第

3 项采用与式 (5) 相同的 L_{21} 范数作为正则化项, 同样用于在多模态脑成像定量特征中联合筛选出与风险 SNP 位点相关的少量特征。该研究是基于阿尔茨海默病神经影像学 ADNI 样本来开展的, 使用来自 37 个阿尔茨海默病相关基因的 1 224 个 SNP 来预测 10 项基于体素的形态学测量 VBM 指标, 并依据连锁不平衡区块 LD 对 SNP 进行分组。

1.4 讨论

除脑影像基因组学中的正则化多变量回归模型外, 诸多贝叶斯算法也相继被提出。例如, 受 GSMuRFS^[52]启发, 研究者构建了贝叶斯组稀疏多任务回归 (Bayesian group sparse multitask regression, BGSMTR) 模型, 以识别多变量基因-多变量影像回归关联, 同时纳入 SNP 数据中的组结构信息 (如连锁不平衡区块和基因)。相较于仅提供回归系数点估计的 G-SMuRFS, BGSMTR 支持完整的后验推断 (如回归参数的区间估计)。该模型可视为贝叶斯组稀疏^[53-54]的扩展, 能够处理多变量响应变量并在 SNP 和基因层面进行变量选择。文献[55]构建了贝叶斯广义低秩回归 (Generalized low-rank regression, GLRR) 模型, 用于分析高维成像响应变量与协变量, 通过低秩表示近似高维权重矩阵。该 GLRR 模型进一步发展为贝叶斯纵向低秩回归 L2R2^[56], 以研究遗传因素对纵向成像响应的影响。

1.1~1.3 节介绍的方法统称为多变量回归模型, 本文仅选取典型案例进行说明。这类方法致力于揭示多变量 SNP 数据与成像 QT 数据间的复杂影像基因组学关联, 其共同原理在于均是通过正则化回归模型识别 SNP 与影像 QT 的关联关系。这些模型具有两大显著优势: (1) 回归系数可直接反映基因 SNP 与影像 QT 的关联性, 便于解释; (2) 基于单一模型获得的遗传标记与成像标记无需多重检验校正, 可提升检测效能。然而, 受数据高维性影响, 此类模型存在过拟合风险。为弥补这一缺陷, 研究者通过引入多种正则化形式降低模型复杂度, 并纳入生物学合理结构以减少过拟合风险。例如, 利用 L_1 或 L_{21} 范数实现稀疏性可简化模型复杂度 (如 G-SMuRFS 和 SRRR 模型); 通过组稀疏或组 L_{21} 范数 (如 P-GLAW 和 P-SRRR 模型), 可纳入生物学结构信息 (如连锁不平衡和通路信息)。此外, 低秩约束可作正则化项, 如任务相关纵向稀疏回归 (Task-correlated longitudinal sparse regression, TCLSR) 模

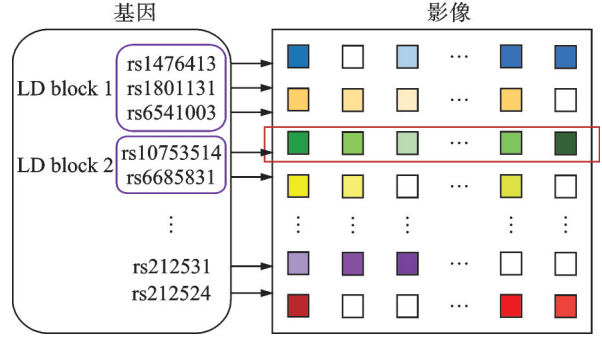


图6 组稀疏多任务回归与特征选择策略^[52]

Fig.6 Group-sparse multi-task regression and feature selection strategy^[52]

型^[57]、时序结构自学习(Temporal structure autolearning, TSAL)模型^[58]和联合投影学习与稀疏回归(Joint projection learning and sparse regression, JPLSR)模型^[59],在有效处理空间或时序相关性的同时降低模型复杂度。

2 双多变量关联

除回归模型外,另一类在脑影像基因组学研究中广泛应用的重要方法是双多变量关联模型,例如稀疏典型相关分析SCCA^[30-32]和平行独立成分分析pICA^[33-34]。与前文讨论的回归模型类似,这些关联模型也通过引入稀疏性来降低模型复杂度、控制过拟合风险,并识别相关生物标志物。设 $\mathbf{X} \in \mathbf{R}^{n \times p}$ 为针对 n 名受试者采集的、包含 p 个变量的遗传数据, $\mathbf{Y} \in \mathbf{R}^{n \times q}$ 为针对同一批 n 名受试者采集的、包含 q 个变量的成像数据。假设 $\mathbf{X}$ 和 $\mathbf{Y}$ 的每一列数据均已进行标准化处理,使其均值为0且方差为1。在脑影像基因组学研究中,应用最为广泛的双多变量关联模型是SCCA及其各类添加不同正则化项的变体模型。这类模型通常可采用以下通用正则化典型相关分析形式进行描述,即

$$\begin{cases} \max_{\mathbf{u}, \mathbf{v}} \mathbf{u}^T \mathbf{X}^T \mathbf{Y} \mathbf{v} - \sum_{i=1}^k \lambda_i R_i(\mathbf{u}, \mathbf{v}) \\ \text{s.t. } \|\mathbf{X} \mathbf{u}\|_2^2 = \|\mathbf{Y} \mathbf{v}\|_2^2 = 1 \end{cases} \quad (11)$$

在脑影像基因组学背景下,其目标在于寻找一个遗传成分 $\mathbf{X} \mathbf{u}$ (即SNP位点的线性组合)和一个影像成分 $\mathbf{Y} \mathbf{v}$ (即成像定量特征QT的线性组合),使得二者相关性(即 $\mathbf{u}^T \mathbf{X}^T \mathbf{Y} \mathbf{v}$, s.t. $\|\mathbf{X} \mathbf{u}\|_2^2 = \|\mathbf{Y} \mathbf{v}\|_2^2 = 1$)在一项或多项正则化约束项 $R_i(\mathbf{u}, \mathbf{v})$ 的条件下达到最大化。例如,传统稀疏典型相关分析模型^[60]通过引入两个 L_1 范数约束项构建: $R_1(\mathbf{u}) = \|\mathbf{u}\|_1$ 和 $R_2(\mathbf{v}) = \|\mathbf{v}\|_1$ 。研究者还可定义其他各类正则化项以实现不同目标,例如将组结构/网络结构或其他脑影像基因组学先验知识纳入模型。以下将通过具体研究案例,探讨采用正则化SCCA策略的若干应用实例。

2.1 多变量基因-多变量影像关联

通过将稀疏性引入多变量基因-多变量影像关联分析,模型能够从高维双变量数据中自动筛选出相关的稀疏SNP位点特征变量和成像定量特征QT特征变量。然而,稀疏典型相关分析SCCA的主要缺陷在于仍未充分考量特征变量间的结构关系,即建模过程中未有效利用大量先验信息。例如,同一连锁不平衡LD区块内的SNP位点可能具有共同特征,而大脑需要多个脑区协同作用才能完成特定功能。因此,在多变量基因-多变量影像关联研究中,为弥补传统SCCA的不足,众多学者通过引入各类先验信息作为正则化项对SCCA模型进行扩展改进^[61-65]。例如,在文献[65]中,构建了结构感知型稀疏典型相关分析(Structure-aware SCCA, S2SCCA)模型,通过在式(10)中添加以下两个组 L_1 范数约束项实现

$$R_1(\mathbf{u}) = \sum_{g \in G_1} \|\mathbf{u}_g\|_2 \quad (12)$$

$$R_2(\mathbf{v}) = \sum_{g \in G_2} \|\mathbf{v}_g\|_2 \quad (13)$$

式(12)中连锁不平衡LD区块被用于构建式(11)中的SNP分组结构 G_1 。式(13)中感兴趣区域ROIs被用于形成体素级成像QT分组结构 G_2 。该研究旨在通过将先验知识作为正则化项,识别APOE SNPs与体素级QT之间的多变量基因-多变量影像关联。

此外,Yan等^[61]提出了一种先验知识引导的稀疏典型相关分析(Knowledge-guided sparse canonical correlation analysis, KG-SCCA),如图7所示。该模型采用组稀疏正则化约束,将连锁不平衡LD区块分

组先验信息的 SNP 位点嵌入到式(11)中,表达式如下

$$R_1(\mathbf{u}) = \sum_{i=1}^g \sqrt{\sum_{j \in G(i)} u_j^2} \leq c_3 \quad (14)$$

式中 c_3 表示阈值。在 L_2 范数约束组内,特征变量尽可能具有相同的权重贡献,也就是说,同一连锁不平衡 LD 区块中的 SNP 位点在关联分析中更有可能被选中。 L_1 范数则通过约束组间稀疏性,筛选出少数具有强相关性的 LD 区块。同时,该模型还将脑功能网络信息作为脑区特征相似性的先验知识引入其中。换句话说,当脑网络中的连接权重较高时,这两个脑区节点具有相似的特征(基因表达高度相关),正则化约束的表达形式如下

$$R_2(\mathbf{v}) = \sum_{(i,j) \in E, i < j} \tau(\mathbf{w}_{ij}) \times \|\mathbf{v}_i - \text{sgn}(\mathbf{w}_{ij}) \mathbf{v}_j\|_2 \leq c_4 \quad (15)$$

式中: c_4 表示阈值; $\mathbf{v}_i$ 和 $\mathbf{v}_j$ 分别代表脑网络上任意两个节点的特征权重; $\text{sgn}(\mathbf{w}_{ij})$ 是 $\mathbf{v}_i$ 和 $\mathbf{v}_j$ 之间相关性的符号函数,当 $\text{sgn}(\mathbf{w}_{ij})$ 为正时, $\mathbf{v}_i$ 和 $\mathbf{v}_j$ 呈正相关,当 $\text{sgn}(\mathbf{w}_{ij})$ 为负时, $\mathbf{v}_i$ 和 $\mathbf{v}_j$ 呈负相关; $\tau(\mathbf{w}_{ij})$ 是 $\mathbf{v}_i$ 和 $\mathbf{v}_j$ 之间的连接强度, $\tau(\mathbf{w}_{ij})$ 值越高,表明脑区变量 $\mathbf{v}_i$ 和 $\mathbf{v}_j$ 越倾向于被同时选中。

此外,部分研究还考虑了基于双多变量多任务先验知识诱导的稀疏典型相关分析 SCCA 模型。例如,文献[66-67]提出了一种多任务稀疏典型相关分析(Multitask SCCA, MTSCCA)模型,用于识别 SNP 数据与多模态影像数据之间的双多变量关联,具体模型如下

$$\begin{cases} \max_{\mathbf{u}, \mathbf{v}} \sum_{j=1}^M \mathbf{u}_j^T \mathbf{X}^T \mathbf{Y}_j \mathbf{v}_j - \lambda_1 \|\mathbf{U}\|_{2,1} - \lambda_2 \|\mathbf{U}\|_{G_{2,1}} - \lambda_3 \|\mathbf{V}\|_{2,1} \\ \text{s.t.} \|\mathbf{X} \mathbf{u}_j\|_2^2 = \|\mathbf{Y}_j \mathbf{v}_j\|_2^2 = 1 \end{cases} \quad (16)$$

式中: $\mathbf{X} \in \mathbb{R}^{n \times p}$ 为 SNP 数据; $\mathbf{Y}_j \in \mathbb{R}^{n \times q} (j \in [1, M])$ 为 M 种模态的影像数据; $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_M]$ 且 $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M]$ 。值得注意的是, $L_{2,1}$ 范数正则化项是一种“组稀疏”正则化项,仅从不同模态中选择少量特征。第 1 个正则化项为 $L_{2,1}$ 范数,用于选择 SNP 特征;第 2 个正则化项为组 $L_{2,1}$ 范数,可在连锁不平衡 LD 区块层面选择 SNP 特征;第 3 个正则化项同样为 $L_{2,1}$ 范数,用于从所有模态中选择影像特征。

在真实的生物医学研究中,提供精确的先验知识是一项艰巨的任务。因此,一旦生物先验知识不可用或不完整,上述扩展形式的稀疏典型相关分析模型可能无效。一般来说,研究采用样本相关性而非先验知识来定义图或网络约束。具体而言,式(11)中使用了 3 种类型的正则化方法:(1) 用于实现平坦稀疏性的 L_1 范数;(2) 用于实现组稀疏性的组 L_1 范数;(3) 图拉普拉斯型范数,用于联合选择图中相连的特征。例如,在文献[68]中,设计了一种基于通用非凸惩罚的稀疏典型相关分析(Generic non-convex penalty based SCCA, GNCSCCA)模型,具体如下

$$R(\mathbf{u}) = \sum_{i=1}^p P_{\lambda, \gamma}(|\mathbf{u}_i|) \quad (17)$$

式中: λ 和 γ 代表非负参数; $P_{\lambda, \gamma}(|\mathbf{u}_i|)$ 表示一个非凸函数。为降低估计偏差,在基于 L_1 范数的稀疏典型相关分析模型中引入了 7 种非凸惩罚项。研究在 ADNI 数据集上开展了一项实证研究,以识别基于体

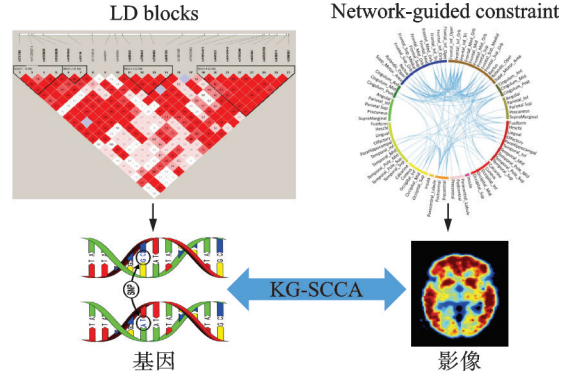


图7 KG-SCCA示意图^[61]

Fig.7 Schematic of KG-SCCA^[61]

素的QTs与来自AD相关基因的163个SNP之间的多SNP-多QT关联。

由于稀疏典型相关分析具备强大的识别双多变量关系并同步进行特征选择的能力,已成为该领域广泛使用的工具。 L_0 范数虽能诱导稀疏性,但属于NP难问题。实际应用中,通常采用 L_1 范数或其变体替代 L_0 范数以实现稀疏诱导。例如,文献[69]提出了截断 L_1 范数惩罚的稀疏典型相关分析(Truncated L_1 norm penalized SCCA, TLP-SCCA)和截断组 Lasso 稀疏典型相关分析(Truncated group Lasso SCCA, TGL-SCCA),二者分别采用以下截断 L_1 范数和截断组 Lasso 方法,即

$$R(\mathbf{u}) = \sum_{i=1}^P J_{\tau}(|\mathbf{u}_i|), \quad J_{\tau}(\mathbf{u}_i) = \min\left(\frac{|\mathbf{u}_i|}{\tau}, 1\right) \quad (18)$$

$$R(\mathbf{u}) = \sum_{k=1}^K J_{\tau}(|\mathbf{G}_k|), \quad J_{\tau}(\mathbf{G}_k) = \min\left(\frac{|\mathbf{G}_k|}{\tau}, 1\right) \quad (19)$$

式中 τ 表示一个可调参数。选择合适的 τ 值, $R(\mathbf{u})$ 能够在 L_0 范数和 L_1 范数之间实现平衡。需要注意的是, $\mathbf{G}_k$ 表示第 k 组($k \in [1, K]$)的一个子集, $\mathbf{u}$ 表示所有 $\mathbf{G}_k$ 的拼接(串联)结果。

此外,受 GraphNet^[70]的启发,Du等^[62]提出了一种基于绝对值的图网络稀疏典型相关分析(Absolute value-based GraphNet SCCA, AGNSCCA),该方法将 GraphNet 正则化的扩展形式引入到 SCCA 模型中。AGN 正则化的形式可描述为

$$R_1(\mathbf{u}) = |\mathbf{u}|^T L_1 \mathbf{u} + \beta_1 \|\mathbf{u}\|_1 \quad (20)$$

$$R_2(\mathbf{v}) = |\mathbf{v}|^T L_2 \mathbf{v} + \beta_2 \|\mathbf{v}\|_1 \quad (21)$$

式中 L_1 和 L_2 分别表示 $\mathbf{X}$ 和 $\mathbf{Y}$ 相关矩阵的拉普拉斯矩阵。应当指出的是,此处采用数据驱动的相关性作为图约束,以便能够共同选择出具有相关性的特征。通过加入绝对值运算,正相关和负相关的特征均可被联合选择出来。

此外,由样本相关性驱动的 SCCA 模型主要致力于探究单时间点上单 SNP 数据与单一成像模式数据之间的关联。幸运的是,这些模型已得到扩展,可聚焦于纵向成像数据。例如,为识别与纵向表型标志物相关的遗传关联,Hao等^[71]设计了一种时间约束组稀疏典型相关分析(Temporally constrained group SCCA, TGSCCA)框架,其模型构建如下

$$\begin{cases} \max_{\mathbf{u}, \mathbf{v}} \sum_{t=1}^T \mathbf{u}^T \mathbf{X}^T \mathbf{Y}_t \mathbf{v}_t - \lambda_1 \|\mathbf{u}\|_1 - \lambda_2 \|\mathbf{V}\|_{2,1} - \lambda_3 \sum_{t=1}^{T-1} \|\mathbf{v}_{t+1} - \mathbf{v}_t\|_1 \\ \text{s.t.} \|\mathbf{X}\mathbf{u}\|_2^2 = \|\mathbf{Y}_t \mathbf{v}_t\|_2^2 = 1 \end{cases} \quad (22)$$

式中 $\mathbf{u}$ 和 $\mathbf{v}_t$ 分别表示在时间点 t 上衡量 SNP 位点和影像表型感兴趣区域 RoI 贡献的权重向量; $\mathbf{v}_{t+1}$ 为相邻时间点的权重向量; λ_1 、 λ_2 和 λ_3 为3个正则化参数。可以看出,融合 Lasso 正则化能够约束相邻时间点上两个连续典型权重向量之间的差距,使其保持较小,从而使权重向量变得平滑,能够共同选择相邻特征。

2.2 讨论

本节将讨论其他双多变量关联模型。Fang等^[72]设计了一种贪婪投影距离相关(Greedy projected distance correlation, G-PDC)策略,用于检验基因与感兴趣区域 RoI 之间的成对关联,其中每个基因和 RoI 分别包含多个 SNPs 和体素。距离相关用于衡量两个随机向量(如基因与 RoI)之间的统计依赖性,可对其间的非线性关系进行建模。投影距离相关则倾向于基于距离相关来衡量条件依赖性^[73]。通过提供一对基因-RoI 组合,可检验其独立性,同时控制其他所有 SNP 和体素的影响。Hao等^[74]提出了一

种包含三路稀疏典型相关(Three-way SCCA, T-SCCA)的分析策略,用于探究遗传标记、影像QTs和临床评分之间的内在关联。Hu等^[75]通过将距离相关融入SCCA模型,设计了一种距离典型相关分析(Distance CCA, DCCA)算法。该算法识别出具有最高距离相关的一组原始SNP和一组原始影像QTs,从而减轻多重检验校正的负担。Wang等^[76]提出了一种多模态判别稀疏典型相关分析(Multi-modality discriminant SCCA, MD-SCCA)方法,将有价值的判别相似性信息融入SCCA模型,以改善学习效果。具体而言,首先通过稀疏表示获取类内受试者之间的判别相似性信息。然后,通过强制实施判别相似性信息构建判别稀疏典型相关分析(Discriminant SCCA, D-SCCA)算法。最后,采用MD-SCCA方法全面探究不同受试者不同模态之间的关系。为考虑原始数据潜在的复杂多子空间结构,Wang等^[77]在关联分析前利用反映数据相似性结构的自表达性对原始输入进行重建。具体而言,首先通过稀疏表示将类内相似性信息应用于自表达网络的构建。然后,采用融合方法将来自多模态脑表型的自表达网络迭代融合为一个网络。最后,为构建和使用融合自表达网络提供了一种实用解决方案,以便利用 L_1 范数挖掘单模态表型与基因型之间的关联,同时利用 L_{21} 范数探究多模态表型与基因型之间的关联。此外,深度学习技术在生物学和医学领域的驱动问题中取得了巨大成功。例如,Wang等^[78-79]提出了一种新颖的深度自重构稀疏典型相关分析DS-SCCA方法,用于识别与功能连接表型标记相关的遗传关联。他们通过实现遗传变异(即SNP)与脑网络(即QT)之间的关系,重点关注从fMRI数据中得出的脑网络特征和连接特征(即连接组)的识别。此外,该工作的主要贡献是首次尝试发现遗传因素如何影响脑连接。

与前文所述的回归模型类似,这些关联模型大多基于正则化SCCA,也具有稀疏性,以降低模型复杂度和过拟合风险,并识别相关生物标志物。在这些SCCA模型中,采用 L_1 或 L_{21} 范数进行特征选择,采用组 L_1 或 L_{21} 范数在组水平上进行特征选择,并利用图拉普拉斯矩阵进行图引导学习。值得注意的是, L_{21} 范数通常包含在多模态和纵向SCCA方法中,以跨模态或时间点选择特征,而融合Lasso或融合相对 L_{21} 范数则常用于沿时间维度平滑相邻权重。本文采用这些策略的研究来识别复杂的多SNP-多QT关联,涵盖以下内容:(1)基础SCCA模型,即S2SCCA和KG-SCCA;(2)增强型SCCA模型,即采用通用非凸惩罚的SCCA框架(GNC-SCCA)、TLP-SCCA、TGL-SCCA、基于绝对值的图网络稀疏典型相关分析(AGN-SCCA)以及FDR-corrected SCCA^[80];(3)多模态和纵向SCCA模型,即MTSCCA和TGSCCA;(4)其他双多变量关联模型,即T-SCCA、MD-SCCA、FSN-SCCA、FSN-GSCCA以及DS-SCCA。

3 结果预测

在影像基因组学中,如何整合脑影像与基因组学数据以预测感兴趣的结果(如损伤评分、疾病阶段和进展状态)也是一个颇具吸引力的研究课题。目前,大多数方法通常采用传统机器学习方法或开发新型学习模型,将影像与基因组学数据相结合以进行结果预测^[52, 81-85]。例如,Dukart等^[81]研究了多模态影像(FDG-PET、MRI和Amyloid-PET)、神经心理学和遗传数据作为潜在生物标志物的作用,以识别未来可能罹患AD的轻度认知障碍(Mild cognitive impairment, MCI)患者。具体而言,首先利用上述不同数据模态的组合构建朴素贝叶斯分类器,以区分AD患者和临床正常(Clinically normal, CN)受试者。然后,将学习到的分类器应用于MCI队列,以预测AD转化状态。相关实验结果表明,利用FDG-PET数据可获得76%的准确率,而利用多模态影像和遗传数据则可获得87%的准确率^[81]。文献[82]创建了一种复合影像遗传评分,以预测MCI向AD的转化。在影像方面,首先采用一种非线性模式识别方法来识别与AD相关的体积区域^[83]。然后,通过对这些区域的影像测量值应用非线性支持向量机(Support vector machine, SVM),为每个个体获得影像评分。在基因组学方面,该技术利用线性SVM对AD

患者和临床正常受试者进行分类,从而为每个受试者得出一个多基因AD相关遗传评分。最后,该技术创建了一种复合影像遗传评分,作为影像评分和遗传评分的加权和,相关结果验证了所提出的复合评分可有效提高预测准确率。Peng等^[84]构建了一种结构化稀疏核学习(Structured sparse kernel learning, SSKL)模型,其中,每个特征均由一个核表示,并采用模态信息对核进行分组,以便在特征和组水平上选择变量。

为了理解从遗传学到脑结构和功能,再到认知、行为和诊断结果的生物学通路,许多研究已探索了基因组学、影像和结果之间的关联,并且目前已有一些基于关联分析的结果预测方法^[77, 86-90]。例如,文献[86]构建了一种判别性稀疏典型相关分析模型,以识别与疾病相关的影像蛋白质组学关联。在没有SNP数据的情况下,分析了从脑脊液和血浆中收集的蛋白质表达数据,并研究了其与影像数量性状QTs和多类别诊断标签(CN、MCI和AD)的关系。文献[87]开发了一种联合学习方法,用于诊断相关的影像基因组学关联分析,该方法结合了稀疏典型相关分析和回归分析SCCAR。在此,设 z 为结果数据,此类方法可描述如下

$$\begin{cases} \max_{u,v} \sum_{i=1}^T \frac{1}{2} \|z - Yv\|_2^2 - u^T X^T Yv + \lambda_1 R_1(u) + \lambda_2 R_1(v) \\ \text{s.t.} \|Xu\|_2^2 = \|Yv\|_2^2 = 1 \end{cases} \quad (23)$$

文献[88]提出了一种多任务协同回归(Multi-task collaborative regression, MT-CoReg)算法,能够获取在影像和基因组两种模态中共同表达且与结果相关的变量。该算法可视为一种联合学习方法,它结合了稀疏典型相关分析SCCA和线性回归,同样利用影像成分来预测结果。受文献[88]的启发,Wang等^[77]构建了所提出的融合自表达网络稀疏典型相关分析(Fusion self-expressive network SCCA, FSN-SCCA)关联模型,并采用知名的多核SVM(Multi-kernel SVM, MK-SVM)^[91]对晚期轻度认知障碍(Late MCI, LMCI)、早期轻度认知障碍(Early MCI, EMCI)、显著记忆退化(Significant memory concern, SMC)、AD和正常对照(Normal control, NC)进行分类。图8展示了所提出的FSN-SCCA关联模型与知名的MK-SVM联合用于结果预测的视觉框架。实验结果表明,在阿尔茨海默病神经影像学倡议ADNI数据集上,该方法在AD与NC、EMCI与LMCI的分类中分别取得了93.76%和73.85%的准确率,对应的曲线下面积(Area under the curve, AUC)值分别为0.95和0.7。

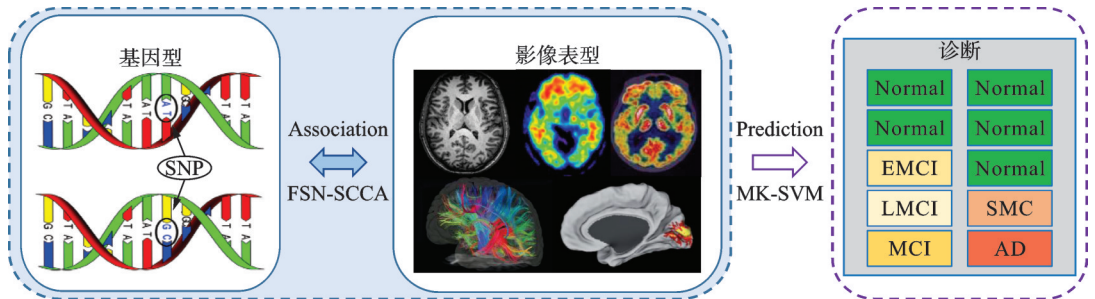


图8 所提出的融合自表达网络关联模型与多核支持向量机联合用于结果预测^[77]

Fig.8 Combination of the proposed FSN-SCCA association model and the MK-SVM for outcome prediction^[77]

需注意的是,大部分方法在临床诊断中通常未考虑与基因型相关的成像表型。众所周知,并非大脑中的所有变化均由遗传效应导致,且通常难以确定哪些成像表型对脑疾病的诊断和预测具有意义。相关结果表明,基于关联分析的结果预测方法有助于指导疾病的解读与预测。

4 结论与展望

影像基因组学作为一门新兴的前沿交叉学科领域,涉及神经科学、成像技术、遗传学、医学、生物统计学、数据挖掘和机器学习等多种科学与研究技术。基因组数据和多模态成像数据(包括不同时间点的纵向脑成像数据)也为影像基因组学研究提供了丰富的数据支撑,使得基因与大脑结构或功能之间的关联致病机制能够通过具有遗传特性的成像内表型得以呈现。机器学习技术作为数据驱动关联分析的强大工具,能够分析易感基因与大脑结构或功能之间的关联,并通过充分挖掘和利用基因与影像等生物标志物数据的内部结构信息,更好地揭示大脑认知行为或相关疾病的机制。

本文回顾了基于机器学习的关联分析算法在影像基因组学研究领域的最新应用,使用了包括PubMed、Scopus和Web of Science在内的3个数据库来筛选所回顾的论文。以“脑影像基因组学(遗传学)”作为主要筛选关键词,设置查询时间为2025年7月20日。大量文献表明,模型检测到的一些关联结果也在生物学和医学领域得到了验证。

本文所介绍的结构化多变量影像遗传关联分析方法均基于某些先验知识,即关联分析是在相对相关的候选基因集或脑区集上进行的。尽管从理论上讲,本文所述方法可应用于全基因组或全脑体素分析,但计算效率较低。为提高高维特征变量基因位点检测的效率,除改进算法本身的效率外,还可通过引入分布式并行计算方法来完成大数据的计算^[92]。因此,有必要进一步开发和构建更高效的算法模型或工作框架,以研究全基因组和全脑特征变量的影像遗传关联。

实际上,基于结构化约束方法的多变量影像遗传研究能够取得良好效果,因为大量先验知识被嵌入到数据分析模型中。例如,作为代表性先验知识之一,LD能够表征SNP之间的简单结构关系。在此基础上,研究人员可以补充先验信息并扩展模型。目前,已有一些工作考虑在模型建立和学习训练中使用具有更多遗传功能的生物特征先验知识,包括基因本体(Gene ontology, GO)、功能注释和通路分析系统,如京都基因与基因组百科全书(Kyoto encyclopedia of genes and genes, KEGG)通路数据库或人类在线孟德尔遗传(Online Mendelian inheritance in man, OMIM)疾病数据库^[93]。因此,如何根据这些先验知识(如神经调节所涉及的机制)设计更适用于实际数据分析应用问题的模型,即实现假设驱动与数据驱动方法的结合,以获得更好的关联结果,仍是当前的研究热点^[94]。

尽管结构化多变量基因-多变量影像关联结果能够解释遗传效应,但同一性状中多个非等位基因之间可能存在交互关系,即异位显性机制尚不十分清楚。目前,已有一些工作研究SNP对影像定量特征QTs的交互作用^[95]。这些方法主要基于遍历成对搜索方法。例如,Hibar等^[96]使用迭代确定性独立筛选(Sure independence screening, SIS)算法来实现并检测与脑区性状显著相关的SNP-SNP交互作用。这些遍历搜索需要耗费大量的计算时间,而一些高效的稀疏模型^[36]有望为多交互作用的异位显性研究提供高效的学习算法。

在结果预测研究中,上述大多数基于监督信息的基因-影像关联研究仅用于研究大脑认知行为或疾病产生的机制,并为疾病诊断和预测提供依据^[77, 86-90]。因此,如何构建一个遗传学、脑成像、临床评分和诊断信息数据的联合关联、回归和分类的多任务统一模型,该模型不仅能够揭示遗传学与脑成像之间的关系,还能够基于生物标志物实现疾病的诊断和预测,这也将成为未来脑影像基因组学研究的发展方向^[77, 86, 97]。

本文工作聚焦于基于机器学习的基因-影像关联研究的主要思路与建模方法。众所周知,深度学习在生物学和医学领域的数据驱动问题中取得了巨大成功。然而,深度学习尚未在脑影像基因组学中得到广泛应用,部分原因在于现有影像和基因组学数据集的样本量有限且维度较高。近期研究已开发出一些基于深度学习的方法,通过结合脑影像基因组学数据来进行结果预测^[87-88, 90, 98]。由于深度学习在

医学图像分析^[99]和多组学研究^[100]中均取得了优异表现,因此开发深度学习方法以解决脑影像基因组学中的紧迫问题是一个具有前景的研究方向。

参考文献:

- [1] THOMPSON P M, MARTIN N G, WRIGHT M J. Imaging genomics[J]. *Current Opinion in Neurology*, 2010, 23(4): 368-373.
- [2] GLAHN D C, THOMPSON P M, BLANGERO J. Neuroimaging endophenotypes: Strategies for finding genes influencing brain structure and function[J]. *Human Brain Mapping*, 2007, 28(6): 488-501.
- [3] WEINER M W, VEITCH D P, ASIEN P S, et al. Recent publications from the Alzheimer's disease neuroimaging initiative: Reviewing progress toward improved ad clinical trials[J]. *Alzheimer's Dementia*, 2017, 13(4): e1-e85.
- [4] THOMPSON P M, JAHANSHAD N, CHING C R K, et al. ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries[J]. *Translational Psychiatry*, 2020, 10: 100.
- [5] SUDLOW C, GALLACHER J, ALLEN N, et al. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age[J]. *PLoS Medicine*, 2015, 12(3): e1001779.
- [6] SHEN L, THOMPSON P M, POTKIN S G, et al. Genetic analysis of quantitative phenotypes in AD and MCI: Imaging, cognition and biomarkers[J]. *Brain Imaging and Behavior*, 2014, 8(2): 183-207.
- [7] 马佳, 吴海锋, 李顺良. 轻度认知障碍分类中全脑功能连接的特征压缩分析[J]. *数据采集与处理*, 2024, 39(4): 967-983.
MA Jia, WU Haifeng, LI Shunliang. Feature compression analysis of whole-brain functional connection in classification of mild cognitive impairment[J]. *Journal of Data Acquisition and Processing*, 2024, 39(4): 967-983.
- [8] VEITCH D P, WEINER M W, AISEN P S, et al. Understanding disease progression and improving Alzheimer's disease clinical trials: Recent highlights from the Alzheimer's disease neuroimaging initiative[J]. *Alzheimer's & Dementia*, 2019, 15(1): 106-152.
- [9] THOMPSON P M, STEIN J L, MEDLAND S E, et al. The ENIGMA consortium large-scale collaborative analyses of neuroimaging and genetic data[J]. *Brain Imaging and Behavior*, 2014, 8(2): 153-182.
- [10] THOMPSON P M, ANDREASSEN O A, ARIAS-VASQUEZ A, et al. ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide[J]. *NeuroImage*, 2017, 145: 389-408.
- [11] BEARDEN C E, THOMPSON P M. Emerging global initiatives in neurogenetics: The enhancing neuroimaging genetics through meta-analysis (ENIGMA) consortium[J]. *Neuron*, 2017, 94(2): 232-236.
- [12] BYCROFT C, FREEMAN C, PETKOVA D, et al. The UK Biobank resource with deep phenotyping and genomic data[J]. *Nature*, 2018, 562(7726): 203-209.
- [13] ALFARO-ALMAGRO F, JENKINSON M, BANGERTER N K, et al. Image processing and quality control for the first 10 000 brain imaging datasets from UK Biobank[J]. *NeuroImage*, 2018, 166: 400-424.
- [14] ELLIOTT L T, SHARP K, ALFARO-ALMAGRO F, et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank[J]. *Nature*, 2018, 562(7726): 210-216.
- [15] GOTTESMAN I I, GOULD T D. The endophenotype concept in psychiatry: Etymology and strategic intentions[J]. *The American Journal of Psychiatry*, 2003, 160(4): 636-645.
- [16] MEYER-LINDENBERG A, WEINBERGER D R. Intermediate phenotypes and genetic mechanisms of psychiatric disorders [J]. *Nature Reviews Neuroscience*, 2006, 7(10): 818-827.
- [17] GE T, SCHUMANN G, FENG J F. Imaging genetics-towards discovery neuroscience[J]. *Quantitative Biology*, 2013, 1(4): 227-245.
- [18] WINKLER A M, KOCHUNOV P, BLANGERO J, et al. Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies[J]. *NeuroImage*, 2010, 53(3): 1135-1146.
- [19] SMITH M, FOX P T, MILLER K L, et al. Correspondence of the brain's functional architecture during activation and rest [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2009, 106(31): 13040-13045.
- [20] TOST H, BILEK E, MEYER-LINDENBERG A. Brain connectivity in psychiatric imaging genetics[J]. *NeuroImage*, 2012,

- 62(4): 2250-2260.
- [21] RUBINOV M, SPORNS O. Complex network measures of brain connectivity: Uses and interpretations[J]. *NeuroImage*, 2010, 52(3): 1059-1069.
 - [22] HARDY J, SINGLETON A. Genomewide association studies and human disease[J]. *The New England Journal of Medicine*, 2009, 360(17): 1759-1768.
 - [23] KLEIN R J, ZEISS C, CHEW E Y, et al. Complement factor H polymorphism in age-related macular degeneration[J]. *Science*, 2005, 308(5720): 385-389.
 - [24] ESSLINGER C, WALTER H, KIRSCH P, et al. Neural mechanisms of a genome-wide supported psychosis variant[J]. *Science*, 2009, 324(5927): 605.
 - [25] LIU J, CALHOUN V D. A review of multivariate analyses in imaging genetics[J]. *Frontiers in Neuroinformatics*, 2014, 8: 29.
 - [26] 郝小可, 李蝉秀, 严景文, 等. 基于统计学习的影像遗传学方法综述[J]. *自动化学报*, 2018, 44 (1): 13-24.
HAO Xiaoke, LI Shanxiu, YAN Jingwen, et al. A review of statistical-learning imaging genetics[J]. *Acta Automatica Sinica*, 2018, 44(10): 13-24.
 - [27] WANG M, SHAO W, HAO X, et al. Machine learning for brain imaging genomics methods: A review[J]. *Machine Intelligence Research*, 2023, 20(1): 57-78.
 - [28] MUFFORD M S, STEIN D J, DALVIE S, et al. Neuroimaging genomics in psychiatry—A translational approach[J]. *Genome Medicine*, 2017, 9(1): 102.
 - [29] SHEN L, THOMPSON P M. Brain imaging genomics: Integrated analysis and machine learning[J]. *Proceedings of the IEEE*, 2020, 108(1): 125-162.
 - [30] WITTEN D M, TIBSHIRANI R, HASTIE T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis[J]. *Biostatistics*, 2009, 10(3): 515-534.
 - [31] CAO K A L, MARTIN P G P, ROBERT-GRANIÉ C, et al. Sparse canonical methods for biological data integration: Application to a cross-platform study[J]. *BMC Bioinformatics*, 2009, 10: 34.
 - [32] CHI E C, ALLEN G I, ZHOU H, et al. Imaging genetics via sparse canonical correlation analysis[C]//*Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*. San Francisco, USA: IEEE, 2013: 740-743.
 - [33] LIU J Y, DEMIRCI O, CALHOUN V D. A parallel independent component analysis approach to investigate genomic influence on brain function[J]. *IEEE Signal Processing Letters*, 2008, 15: 413-416.
 - [34] CALHOUN V D, LIU J Y, ADALIMATH T. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data[J]. *NeuroImage*, 2009, 45(1): S163-S172.
 - [35] KOHANNIM O, HIBAR D P, STEIN J L, et al. Discovery and replication of gene influences on brain structure using LASSO regression[J]. *Frontiers in Neuroscience*, 2012, 6: 115.
 - [36] YANG T, WANG J, SUN, Q, et al. Detecting genetic risk factors for Alzheimer's disease in whole genome sequence data via Lasso screening[C]//*Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*. Brooklyn, USA: IEEE, 2015: 985-989.
 - [37] SILVER M, MONTANA G. Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps[J]. *Statistical Applications in Genetics and Molecular Biology*, 2012, 11(1): 7.
 - [38] YUAN M, LIN Y. Model selection and estimation in regression with grouped variables[J]. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2006, 68(1): 49-67.
 - [39] SILVER M, CHEN P, LI R Y, et al. Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two Asian cohorts[J]. *PLoS Genetics*, 2013, 9(11): e1003939.
 - [40] BARRETT J C, FRY B, MALLER J, et al. Haploview: Analysis and visualization of LD and haplotype maps[J]. *Bioinformatics*, 2005, 21(2): 263-265.
 - [41] HAO X K, YU J T, ZHANG D Q. Identifying genetic associations with MRI-derived measures via tree-guided sparse learning [C]//*Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Boston, USA: Springer, 2014: 757-764.
 - [42] HAO X K, YAO X H, RISACHER S L, et al. Identifying candidate genetic associations with MRI-derived AD-related ROI via

- tree-guided sparse learning[J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019, 16(6): 1986-1996.
- [43] WANG J, YE J P. Multi-layer feature reduction for tree structured group lasso via hierarchical projection[C]//*Proceedings of International Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada: [s.n.], 2015: 1279-1287.
- [44] SHEN L, KIM S, RISACHER S L, 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[J]. *NeuroImage*, 2010, 53(3): 1051-1063.
- [45] HAO X K, YAN J W, YAO X H, et al. Diagnosis-guided method for identifying multi-modality neuroimaging biomarkers associated with genetic risk factors in Alzheimer's disease[C]//*Proceedings of Pacific Symposium on Biocomputing(PSB)*. Kohala Coast, USA: [s.n.], 2016: 108-119.
- [46] HAO X K, YAO X H, YAN J W, et al. Identifying multimodal intermediate phenotypes between genetic risk factors and disease status in Alzheimer's disease[J]. *Neuroinformatics*, 2016, 14(4): 439-452.
- [47] WANG M L, HAO X K, HUANG J, et al. Discovering network phenotype between genetic risk factors and disease status via diagnosis-aligned multi-modality regression method in Alzheimer's disease[J]. *Bioinformatics*, 2019, 35(11): 1948-1957.
- [48] TIBSHIRANI R. Regression shrinkage and selection via the lasso: A retrospective[J]. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2011, 73(3): 273-282.
- [49] VOUNOU M, NICHOLS T E, MONTANA G, et al. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach[J]. *NeuroImage*, 2010, 53(3): 1147-1159.
- [50] VOUNOU M, JANOUSOVA E, WOLZ R, et al. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease[J]. *NeuroImage*, 2012, 60(1): 700-716.
- [51] SILVER M, JANOUSOVA E, HUA X, et al. Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression[J]. *NeuroImage*, 2012, 63(3): 1681-1694.
- [52] WANG H, NIE F P, HUANG S, et al. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort[J]. *Bioinformatics*, 2012, 28(2): 229-237.
- [53] PARK T, CASELLA G. The Bayesian lasso[J]. *Journal of the American Statistical Association*, 2008, 103(482): 681-686.
- [54] CASELLA G, GHOSH M, GILL J, et al. Penalized regression, standard errors, and Bayesian lassos[J]. *Bayesian Analysis*, 2010, 5(2): 369-411.
- [55] ZHU H T, KHONDKER Z, LU Z H, et al. Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers[J]. *Journal of the American Statistical Association*, 2014, 109(507): 977-990.
- [56] LU Z H, KHONDKER Z, IBRAHIM J G, et al. Bayesian longitudinal low-rank regression models for imaging genetic data from longitudinal studies[J]. *NeuroImage*, 2017, 149: 305-322.
- [57] WANG H, NIE F P, HUANG H, et al. From phenotype to genotype: An association study of longitudinal phenotypic markers to Alzheimer's disease relevant SNPs[J]. *Bioinformatics*, 2012, 28(18): i619-i625.
- [58] WANG X Q, YAN J W, YAO X H, et al. Longitudinal genotype-phenotype association study through temporal structure auto-learning predictive model[J]. *Journal of Computational Biology*, 2018, 25(7): 809-824.
- [59] ZHOU T, THUNG K H, LIU M X. Brain-wide genome-wide association study for Alzheimer's disease via joint projection learning and sparse regression model[J]. *IEEE Transactions on Biomedical Engineering*, 2019, 66(1): 165-175.
- [60] WITTEN D M, TIBSHIRANI R, HASTIE T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis[J]. *Biostatistics*, 2009, 10(3): 515-534.
- [61] YAN J W, DU L, KIM S, et al. Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm[J]. *Bioinformatics*, 2014, 30(17): i564-i571.
- [62] DU L, HUANG H, YAN J W, et al. Structured sparse canonical correlation analysis for brain imaging genetics: An improved GraphNet method[J]. *Bioinformatics*, 2016, 32(10): 1544-1551.
- [63] LIN D D, CALHOUN V D, WANG Y P. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis[J]. *Medical Image Analysis*, 2014, 18(6): 891-902.
- [64] FANG J, LIN D D, SCHULZ S C, et al. Joint sparse canonical correlation analysis for detecting differential imaging genetics modules[J]. *Bioinformatics*, 2016, 32(22): 3480-3488.
- [65] DU L, YAN J W, KIM S, et al. A novel structure-aware sparse learning algorithm for brain imaging genetics[C]//*Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Boston, USA:

- Springer, 2014: 329-336.
- [66] DU L, LIU K F, YAO X H, et al. Fast multi-task SCCA learning with feature selection for multi-modal brain imaging genetics [C]//Proceedings of IEEE International Conference on Bioinformatics and Biomedicine(BIBM). Madrid, Spain: IEEE, 2018: 356-361.
 - [67] DU L, LIU K F, YAO X H, et al. Multi-task sparse canonical correlation analysis with application to multi-modal brain imaging genetics[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(1): 227-239.
 - [68] DU L, LIU K F, YAO X H, et al. Pattern discovery in brain imaging genetics via SCCA modeling with a generic non-convex penalty[J]. Scientific Reports, 2017, 7(1): 14052.
 - [69] DU L, LIU K F, ZHANG T, et al. A novel SCCA approach via truncated-norm and truncated group lasso for brain imaging genetics[J]. Bioinformatics, 2018, 34(2): 278-285.
 - [70] GROSENICK L, KLINGENBERG B, KATOVICH K, et al. Interpretable whole-brain prediction analysis with GraphNet[J]. NeuroImage, 2013, 72: 304-321.
 - [71] HAO X K, LI C K, YAN J W, et al. Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis[J]. Bioinformatics, 2017, 33(14): i341-i349.
 - [72] FANG J, XU C, ZILLE P. Fast and accurate detection of complex imaging genetics associations based on greedy projected distance correlation[J]. IEEE Transactions on Medical Imaging, 2018, 37(4): 860-870.
 - [73] FAN J Q, FENG Y, XIA L. A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models[J]. Journal of Econometrics, 2020, 218(1): 119-139.
 - [74] HAO X K, LI C X, DU L et al. Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease[J]. Scientific Reports, 2017, 7: 44272.
 - [75] HU W X, ZHANG A Y, CAI B, et al. Distance canonical correlation analysis with application to an imaging-genetic study[J]. Journal of Medical Imaging, 2019, 6(2): 026501.
 - [76] WANG M L, SHAO W, HAO X K, et al. Identify consistent cross-modality imaging genetic patterns via discriminant sparse canonical correlation analysis[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(4): 1549-1561.
 - [77] WANG M L, SHAO W, HAO X K, et al. Identify complex imaging genetic patterns via fusion self-expressive network analysis[J]. IEEE Transactions on Medical Imaging, 2021, 40(6): 1673-1686.
 - [78] WANG M L, SHAO W, HUANG S, et al. Deep self-reconstruction sparse canonical correlation analysis for brain imaging genetics[C]//Proceedings of IEEE International Symposium on Biomedical Imaging(ISBI). Nice, France: IEEE, 2021: 1790-1793.
 - [79] WANG M L, SHAO W, HAO X K, et al. Identify connectome between genotypes and brain network phenotypes via deep self-reconstruction sparse canonical correlation analysis[J]. Bioinformatics, 2022, 38(8): 2323-2332.
 - [80] GOSSMANN A, ZILLE P, CALHOUN V, et al. FDR-corrected sparse canonical correlation analysis with applications to imaging genomics[J]. IEEE Transactions on Medical Imaging, 2018, 37(8): 1761-1774.
 - [81] DUKART J, SAMBATARO F, BERTOLINO A, et al. Accurate prediction of conversion to Alzheimer's disease using imaging, genetic, and neuropsychological biomarkers[J]. Journal of Alzheimer's Disease, 2016, 49(4): 1143-1159.
 - [82] FILIPOVYCH R, GAONKAR B, DAVATZIKOS C. A composite multivariate polygenic and neuroimaging score for prediction of conversion to Alzheimer's disease[C]//Proceedings of IEEE International Workshop on Pattern Recognition in NeuroImaging (PRNI). London, UK: IEEE, 2012: 105-108.
 - [83] FAN Y, SHEN D G, GUR R C, et al. COMPARE: Classification of morphological patterns using adaptive regional elements [J]. IEEE Transactions on Medical Imaging, 2007, 26(1): 93-105.
 - [84] PENG J L, AN L, ZHU X F, et al. Structured sparse kernel learning for imaging genetics based Alzheimer's disease diagnosis [C]//Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI). Athens, Greece: Springer, 2016: 70-78.
 - [85] SINGANAMALLI A, WANG H B, MADABHUSHI A, et al. Cascaded multi-view canonical correlation (CaMCCo) for early diagnosis of Alzheimer's disease via fusion of clinical, imaging and omic features[J]. Scientific Reports, 2017, 7(1): 8137.
 - [86] WANG M L, SHAO W, HUANG S, et al. Hypergraph-regularized multimodal learning by graph diffusion for imaging

- genetics based Alzheimer's disease diagnosis[J]. *Medical Image Analysis*, 2023, 89: 102883.
- [87] BI X, WANG Y, LUO S, et al. Hypergraph structural information aggregation generative adversarial networks for diagnosis and pathogenetic factors identification of Alzheimer's disease with imaging genetic data[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, 35 (6): 7420-7434.
- [88] BI X, HUANG Y, YANG Z, et al. Structure mapping generative adversarial network for multi-view information mapping pattern mining[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46 (4): 2252-2266.
- [89] WANG M L, SUN L, SHAO W, et al. Discovering differential imaging genetic modules via multimodal fusion-based hypergraph transductive learning in Alzheimer's disease diagnosis[J]. *IEEE Transactions on Medical Imaging*, 2025. DOI: 10.1109/TMI.2025.3578601.
- [90] WANG M L, SHAO W, ZHANG D Q, et al. Deep ring-wise block network for joint association analysis and Alzheimer's disease diagnosis with interpretability[J]. *IEEE Transactions on Computational Biology and Bioinformatics*, 2025. DOI: 10.1109/TCBBIO.2025.3583565.
- [91] ZHANG D Q, WANG Y P, ZHOU L P, et al. Multimodal classification of Alzheimer's disease and mild cognitive impairment [J]. *NeuroImage*, 2011, 55(3): 856-867.
- [92] WANG Y, GOH W, WONG L, et al. Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes[J]. *BMC Bioinformatics*, 2013, 14(Suppl 16): S6.
- [93] YAO X H, YAN J W, KIM S, et al. Two-dimensional enrichment analysis for mining high-level imaging genetic associations [C]//*Proceedings of International Conference on Brain Informatics and Health (BIH)*. London, UK: Springer, 2015: 115-124.
- [94] HUYS Q J M, MAIA T V, FRANK M J. Computational psychiatry as a bridge from neuroscience to clinical applications[J]. *Nature Neuroscience*, 2016, 19(3): 404-413.
- [95] BIRNBAUM R, WEINBERGER D R. Functional neuroimaging and schizophrenia: A view towards effective connectivity modeling and polygenic risk[J]. *Dialogues in Clinical Neuroscience*, 2013, 15(3): 279-289.
- [96] HIBAR D P, STEIN J L, JAHANSHAD, et al. Exhaustive search of the SNP-SNP interactome identifies epistatic effects on brain volume in two cohorts[C]//*Proceedings of the 16th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Nagoya, Japan: Springer, 2013: 600-607.
- [97] GROSS S M, TIBSHIRANI R. Collaborative regression[J]. *Biostatistics*, 2015, 16(2): 326-338.
- [98] ZHOU T, THUNG K H, ZHU X, et al. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis[J]. *Human Brain Mapping*, 2019, 40(3): 1001-1016.
- [99] CHEN L, CHEN Q, REN K, et al. Explainable spatio-temporal graph evolution learning with applications to dynamic brain network analysis during development[J]. *NeuroImage*, 2024, 298: 120771.
- [100] DU L, ZHANG J, ZHAO Y, et al. inMTSCCA: An integrated multi-task sparse canonical correlation analysis for multi-omic brain imaging genetics[J]. *Genomics, Proteomics & Bioinformatics*, 2023, 21(2): 396-412.

作者简介:



汪美玲(1988-),女,教授,研究方向:机器学习、脑影像基因组学,E-mail:mely@nj-upt.edu.cn。



刘青山(1975-),男,教授,研究方向:计算机视觉、人工智能等,E-mail:qslu@nj-upt.edu.cn。



张道强(1978-),通信作者,男,教授,研究方向:医学影像分析、人工智能等,E-mail:dqzhang@nuaa.edu.cn。

(编辑:张黄群)