

基于统计学习的影像遗传学方法综述

郝小可¹ 李蝉秀¹ 严景文² 沈理² 张道强¹

摘 要 近年来随着多模态神经影像技术和基因检测技术的发展, 影像遗传学这一交叉学科的研究能够运用脑影像技术将人类大脑的结构与功能作为表型来评价基因对个体的影响, 使得人们可以在脑的宏观结构上以更客观的测量手段理解基因对行为或精神疾病的影响. 而统计学习方法作为基于数据驱动的关联分析强有力工具, 能够充分利用生物标志数据内在的结构信息构建模型来分析易感基因与大脑结构或者功能的相关性, 从而更好地揭示脑认知行为或者相关疾病的产生机制. 本文首先简要介绍了影像遗传学的研究背景和基本原理, 然后回顾了单变量方法在影像遗传学研究中的应用, 随后对基于多变量统计学习的基因-影像关联的研究思路和建模方法进行了归纳总结, 最后对遗传影像学的未来研究发展方向进行了分析和展望.

关键词 影像遗传学, 统计学习, 结构化稀疏学习, 多变量分析, 关联分析

引用格式 郝小可, 李蝉秀, 严景文, 沈理, 张道强. 基于统计学习的影像遗传学方法综述. 自动化学报, 2018, 44(1): 13–24

DOI 10.16383/j.aas.2018.c160696

A Review of Statistical-learning Imaging Genetics

HAO Xiao-Ke¹ LI Chan-Xiu¹ YAN Jing-Wen² SHEN Li² ZHANG Dao-Qiang¹

Abstract The past decade has witnessed the increasing development of multimodal neuroimaging and genomic techniques. Imaging genetics, an interdisciplinary field, aims to evaluate and characterize genetic variants in individuals that influence phenotypic measures derived from structural and functional brain images. This strategy is able to reveal the complex mechanisms via macroscopic intermediates from genetic level to cognition and psychiatric disorders in humans. On the other hand, statistical learning methods, as a powerful tool in the data-driven based association study, can make full use of priori-knowledge (inter correlated structure information among imaging and genetic data) for correlation modelling. Therefore, the association study can address the correlations between risk gene and brain structure or function, so as to help explore a better mechanistic understanding of behaviors or disordered brain functions. This paper firstly reviews the related background and fundamental work in imaging genetics and then shows the univariate statistical learning approaches for correlation analysis. Subsequently, it summarizes the main idea and modeling in gene-imaging association studies based on multivariate statistical learning. Finally, this paper presents some prospects of future work.

Key words Imaging genetics, statistical learning, structured sparse learning, multivariate analysis, association analysis

Citation Hao Xiao-Ke, Li Chan-Xiu, Yan Jing-Wen, Shen Li, Zhang Dao-Qiang. A review of statistical-learning imaging genetics. *Acta Automatica Sinica*, 2018, 44(1): 13–24

近年来, 神经影像学伴随着认知神经科学的发展为人脑工作机制的研究带来了新的活力. 同时随着无创式脑成像技术的发展, 研究者们希望能够从脑结构和脑功能的层次来研究与情绪加工相关的脑活动影响, 从而探索神经系统疾病易感性个体差异的神经基础. 其中, 常用的相关脑成像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)、正电子发射断层扫描成像 (Positron emission tomography, PET). 与此同时, 随着遗传学技术的发展, 研究者们可以从更精细的分子水平 (例如单核苷酸多态性 (Single nucleotide polymorphism, SNP)) 来寻找神经系统疾病和精神疾病相关的遗传标记.

在神经影像学 and 分子遗传学的基础之上, Hariri 等提出了影像遗传学 (Imaging genetics 或 Imaging genomics) 这一概念, 即结合多模态神经影像学 and 遗传学方法, 检测脑结构及与神经疾病、认知和情绪调节等行为相关脑功能的遗传变异^[1–3].

其运用脑影像技术将脑的结构与功能作为表型来评价基因对个体的影响, 探讨基因是如何影响大脑的神经结构和功能, 以及由此导致的神经系统病理. 研究遗传与大脑结构和功能的相关性, 在“基

收稿日期 2016-09-30 录用日期 2017-04-10
Manuscript received September 30, 2016; accepted April 10, 2017

国家自然科学基金 (61422204, 61473149, 61732006) 资助
Supported by National Natural Science Foundation of China (61422204, 61473149, 61732006)

本文责任编辑 朱朝喆

Recommended by Associate Editor ZHU Chao-Zhe

1. 南京航空航天大学计算机科学与技术学院 南京 211106 中国
2. 印第安纳大学医学院 印第安纳波利斯 46202 美国

1. School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
2. School of Medicine, Indiana University, Indianapolis, IN 46202, USA

因与脑”之间架起一座看得见的桥梁^[4-6],可以更好地揭示神经精神疾病的发病机制. 影像遗传学这种工具同时还可以识别出某种脑疾病的生物学指标或其内表型,为预测和诊断疾病提供了更精确的方法. 具体来说,由于 SNP 是在基因组水平上单个核苷酸变异引起的 DNA 序列多态性,在一定程度上反映了个体的遗传特性,因此,研究者大多考虑将 SNP 作为关联分析的基因型数据. 在内表型数据获取中,研究者大多采用临床上广泛使用的 MRI 脑影像数据进行分析: sMRI 作为度量大脑结构组织的成像技术,能够量化分析形态学(如灰质体积)的异常; fMRI 作为血氧水平依赖功能成像技术,无论静息态还是任务态,都能够反映不同脑区的激活程度,从而产生明显的信号差异. 基于不同模态脑成像技术,目前影像遗传学主要关注基因 SNP 与脑结构、功能、连接关联分析的相关研究^[7-10].

早期的影像遗传学是单变量成对的统计分析方法,即通过多次检验,发现 SNP 或者基因与复杂疾病或可测的数量性状(Quantitative trait, QT)的关联性研究方法. 而全基因组关联研究(Genome-wide association study, GWAS)正是利用全基因组高通量测序技术,对研究对象的基因组中序列变异进行分型,并利用生物统计学和生物信息学的方法,最终筛选出具有显著性的 SNP^[11]. 自从 2005 年 Science 上发表的第一篇有关年龄相关性视网膜黄斑变性(Age-related macular degeneration) GWAS 研究论文^[12]以来,该方法也被用在精神疾病的分析上^[13]. GWAS 在影像遗传学的研究中发挥了极大的作用,但是也存在一些问题,比如,严格的多重校正,使得许多微小效应的变异无法通过校正水平. 其次, GWAS 仅仅能得到遗传变异跟性状之间的单个关联程度,并不能很好地解释其中的复杂机制.

近年来,随着统计学习在学术界和工业界迅速发展,许多领域已经尝试利用这些数据分析工具来解决本领域的一些问题. 而在影像遗传学的关联分析中,相对于单变量统计分析,基于多变量的统计学习技术的应用最为广泛,同时也取得了非常理想的效果. 国际上,一些学者也撰写了影像遗传学的相关方法综述文献: 1) Medland 等针对使用传统的单变量统计模型处理大规模全基因组-全脑影像关联分析提出了所面临的问题和挑战,回顾了研究者在不同中心数据库(其中包括 ENIGMA¹、IMAGEN²、IMAGENMEND³以及 ADNI⁴等)的研究成果^[14]; 2) Liu 等主要对独立成

分分析(ICA)等其他多变量方法在影像遗传学中的应用进行了归纳和总结^[15]; 3) Thompson 等在综述中重点回顾了基因与大脑结构连接(DTI)与功能网络(静息态 fMRI)之间的相关分析工作^[16]. 本文在以上综述工作的基础上,首先对基于统计学习的遗传-影像关联研究进展进行回顾,如图 1 所示,其中包括单变量和多变量统计学习方法;本文重点关注基于结构化的多变量分析建模思路和算法框架,即通过生物学过程以及医学领域知识(如代谢通路/网络、多模态融合、诊断信息等)诱导的方法获得更好的关联性能和生物解释;最后,对遗传影像学中一些待解决的问题以及未来研究发展方向进行了展望.

1 单变量分析方法

单基因变量统计分析中最常见的方法是设立实验组和对照组进行皮尔森卡方检验(Pearson's chi-squared test)作为等位基因检测方法,即通过分析各种病症的一组病人和一组正常对照者的相应基因组位点之间是否有统计差异来确认该位点是否是致病基因的. 基于单变量统计方法的基因-影像关联分析可以使用线性回归(Linear regression)和方差分析(Analysis of variance)模型作为等位基因的关联分析方法^[17]. 多次单变量模型,假设基因特征维数为 p ,影像特征维数为 q ,则需要拟合 $p \times q$ 个线性回归模型($y_j = \beta_{jk}x_k$),检测所有 $p \times q$ 个零假设(null hypotheses $H_0: \beta_{jk} = 0$),最后对 p 值(p -value)进行排序. 例如,一个较早的经典工作是来自 2009 年 Potkin 等在病例与对照组和影像表现型上进行全基因组 GWAS 关联分析,SNP 对脑区定量表现型的影响可以通过广义的线性模型来计算,该模型由影像表现型、疾病诊断和基因数据共同构建,表达式如下:

$$Y = b_0 + b_1 \cdot \text{SNP} + b_2 \cdot \text{APOE}\epsilon 4 + b_3 \cdot \text{gender} + b_4 \cdot \text{age} + b_5 \cdot \text{diagnosis} + b_6 \cdot \text{SNP} \times \text{diagnosis} + \epsilon \quad (1)$$

其中, Y 表示神经影像某一脑区的 QT, b_i 表示各个变量系数, $\text{SNP} \times \text{diagnosis}$ 表示相互作用的关系. 模型分析得到的显著 p 值即为 SNP 与 QT 相关的检测结果^[18].

在单变量基因-脑影像关联检测中,我们根据研究问题的规模,将其归纳成不同的尺度^[19]: 在基因层面包括 1) 候选基因/SNP^[20-23], 2) 相关生物功能特性通路/网络^[24-26], 3) 全基因组^[18, 27-30]; 相应的在脑影像层面包括 1) 个别感兴趣区域^[18, 20, 24, 27], 2) 包含多个感兴趣区的回路^[21, 25, 28], 3) 全脑^[22-23, 26, 29-30]. 无论是候选基因位点 SNP

¹<http://enigma.ini.usc.edu/>

²<http://www.imagen-europe.com/>

³<http://www.imagemend.eu/>

⁴<http://adni.loni.usc.edu/>

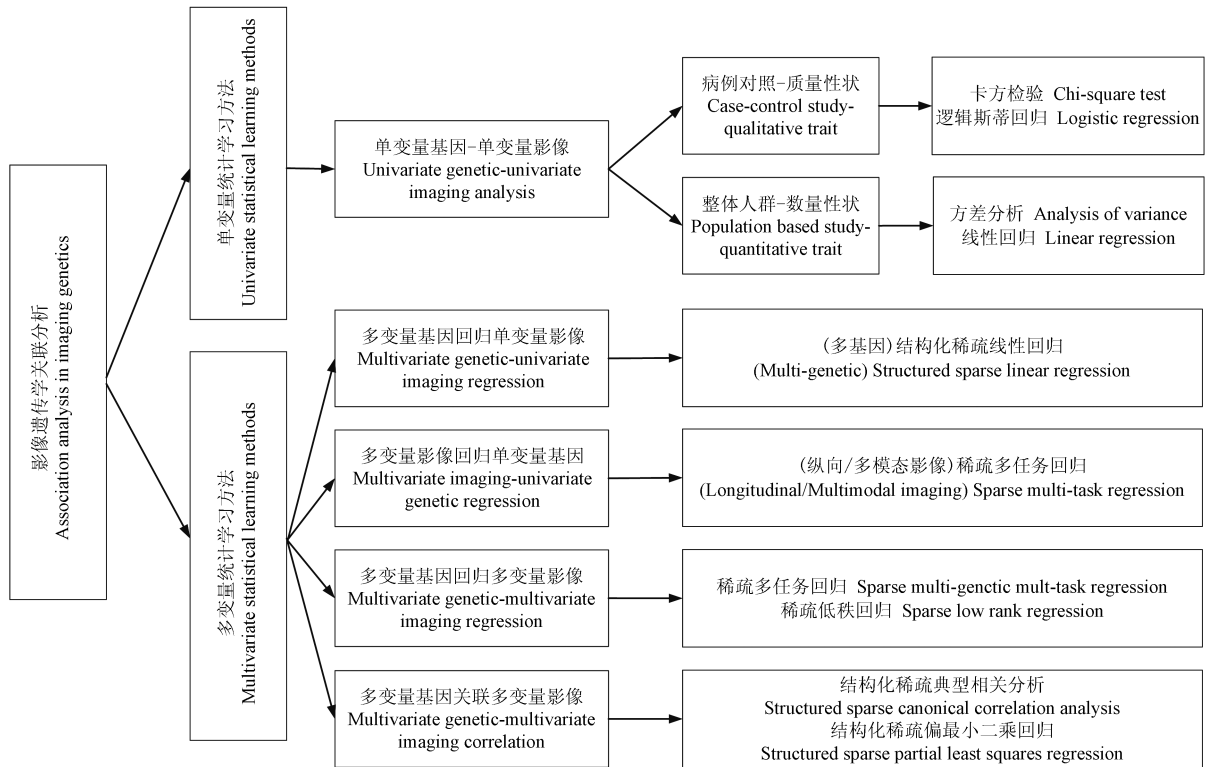


图1 基于统计学习的影像遗传学关联分析方法

Fig. 1 Association analysis in imaging genetics based on statistical learning

与神经影像^[31]、脑脊液^[32]、认知量表得分^[33]等其他任何 QT 关联分析, 还是全基因组与神经影像关联分析^[29], 甚至可以考虑全基因组与更小粒度的体素级别的脑影像之间进行关联分析^[30], 线性回归与方差分析的方法都可以解决不同尺度的影像遗传学研究问题. 在单变量基因影像关联分析研究中, 有些研究者已经发布了相关的统计分析软件, 如 Plink^{5[34]}.

GWAS 遗传统计分析要上百万甚至上千万个 SNP 中发现与疾病表型的关联. 尽管可以利用 Bonferroni 校正来严格地控制显著性^[35-36], 但是这种策略会导致许多微小效应的变异无法通过校正水平, 而多个这样的微小效应变异有可能会共同作用从而对性状产生较大的影响. 单变量分析方法在影像遗传学中的应用具有较为直观的解释性, 能够简单快速地检测出单个 SNP 与单个 QT 之间的关联程度, 但由于数据变量的高维特性而导致的很大数量的多重比较最终使得统计测试结果不具有显著性, 而且上述检验方法基于一个严格的假设, 即基因位点或者影像特征变量之间是统计独立的, 而忽略了变量之间相关性这一重要信息. 因此, 面对单变量方法存在的不足, 在高维特征的基因-影像关联分析这一研究问题中仍然需要在方法学上进行改进和创新.

2 多变量分析方法

继 2010 年 Stein 等提出基于单变量体素级别的全基因组关联分析 (vGWAS)^[30] 之后, Hibar 等提出了一种基于多变量的体素级别全基因组关联分析 (Voxel-wise gene-wide association study, vGeneWAS)^[37-38]. 该方法将一个基因内的所有 SNP 通过主成分回归 (Principal components regression, PCReg) 的方法来解决变量共线性的问题, 首先在 SNP 回归变量集上使用主成分分析 (Principle component analysis, PCA) 获得最大化方差的相互正交因子, 然后对这些正交因子使用标准的偏 F 测试 (Partial F-test). Hibar 等使用与 Stein 等在 2010 年工作中相同的基因和脑影像数据集, 通过 SNP 与体素级别的影像进行关联测试. 实验结果表明, 该方法获得了更好的关联性能, 并且减少了统计测试的次数. 因此, 为了增强基因与性状的关联检测能力, 一些学者和研究人员通过使用多变量方法来解决影像遗传学中多基因或多位点联合效应的关联问题^[15, 39]. 近年来, 基于统计学习的影像遗传学研究备受关注, 很多工作是通过求解目标函数的优化问题来实现检测和识别具有高度关联的基因和影像

⁵<http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml>

特征. 而基因-影像关联问题可以通过多变量基因特征输入 X 、单变量影像表型 QT 输出 Y 的广义线性回归函数进行描述, 模型表达式如下:

$$\min_{\mathbf{w}} f(\mathbf{w}) = L(\mathbf{w}) + \lambda \Omega(\mathbf{w}) \quad (2)$$

其中 $L(\mathbf{w})$ 是损失函数, 通常使用最小平方损失作为回归的损失函数, 即 $L(\mathbf{w}) = \|Y - X\mathbf{w}\|_2^2$. $\Omega(\mathbf{w})$ 为在输入特征变量中嵌入先验知识的正则化项. $\lambda > 0$ 作为正则化参数来平衡损失项与正则化项. \mathbf{w} 为最终学习得到的参数即作为基因与影像的关联权重. 而根据生物医学先验知识, 在数百万 SNP 位点或者数十万的体素脑影像特征中仅仅有很少数量的变量具有高度相关性, 因此高维遗传位点和影像特征的稀疏表示或者稀疏特征选择在生物数据关联分析中具有合理的解释性^[40-43]. 稀疏表示最初在压缩感知 (Compressed sensing) 问题的研究中获得了良好的效果, 其直觉的想法是将大量的稀疏信号用最小数量的线性特征组合来进行编码^[44]. 在统计学习领域, 基于 $L1$ 范数的正则化约束如 Lasso (Least absolute shrinkage and selection operator) 被广泛地用于回归模型中^[45].

$$\Omega_{\text{Lasso}}(\mathbf{w}) = \|\mathbf{w}\|_1 = \sum_i |\mathbf{w}_i| \quad (3)$$

Kohannim 等使用改进的岭回归 (Ridge regression) 模型增强了基因与影像关联的检测能力^[46-47], 即使使用基于 Lasso 的回归模型评估了全基因组与候选脑区体积的关联效应, 发现了多个敏感基因, 且这些基因通过了全基因组显著性测试.

2.1 多变量基因回归单变量影像

基于 $L1$ 范数惩罚约束的回归模型^[48-49] 已经被成功地应用到多变量基因数据分析中, 即识别出与特定脑区高度关联的稀疏 SNP 位点, 从而为处理检测和识别高维基因 SNP 特征选择的小样本回归问题提供了一种普适的技术框架. 然而, 基于 $L1$ 范数的约束并没有充分考虑特征变量之间的结构关系, 因此, 在理论上并不能达到最佳的回归结果. 在考虑了融合 SNP 特征之间的空间结构关系之后, Silver 等提出了使用成组稀疏^[50-52] 的模型来解决影像遗传学问题. 这些关联模型倾向于选择处于同一个组中的 SNP 位点或者倾向于选择相邻的特征变量, 其正则化形式表达如下:

$$\Omega_{\text{groupLasso}}(\mathbf{w}) = \sum_{i=1}^g \sqrt{\sum_{j \in G(i)} \mathbf{w}_j^2} \quad (4)$$

$$\Omega_{\text{fusedLasso}}(\mathbf{w}) = \sum_{i < j} |\mathbf{w}_i - \mathbf{w}_j| \quad (5)$$

其中, 组稀疏约束项 $\Omega_{\text{groupLasso}}(\mathbf{w})$ 中的 \mathbf{w}_j 表示属于 $G(i)$ 组中的所有位点特征, 该目标为控制选择的位点具有聚类的特性. 例如基因的位点之间会产生连锁不平衡 (Linkage disequilibrium, LD) 效应^[53], 即不同基因座位上连锁的 SNP 会非随机地出现在同一个 LD 块中; 这就为基于组稀疏的特征选择模型提供了领域知识, 使得在同一个 LD 组中的 SNP 被同时检测到. 融合 Lasso 的约束项 $\Omega_{\text{fusedLasso}}(\mathbf{w})$ 能够控制相邻位置特征的权重贡献 \mathbf{w}_i 与 \mathbf{w}_j 尽可能相似, 即约束所选出的特征变量具有空间上的连续性. 除了上述考虑 SNP 位点之间平坦的空间关系, 在实际的基因结构中也存在着层次结构关系. 例如, 在某一代谢通路 (Pathway) 中, 特定基因集合共同作用能够在一定程度上影响蛋白的合成以及功能的转化, 而在同一基因下的某些 SNP 位点也具有一定的相关关系 (如 LD). 因此, 充分利用这种层次结构的先验知识进行建模来完成影像遗传学分析往往会降低回归分析中的误差同时能够学习到更具解释意义的特征模式^[54-55], 如图 2 所示, 该模型使用树型结构引导稀疏学习方法识别基因-影像关联. 在构建树型结构时, SNP 位点作为叶子节点, LD 块与基因块作为中间节点, 通路中的全体基因集合作为最终的根节点; 结构树共有 d 层, 每一层有 n_i 个结点, 第 i 层的结点为 $\{G_1^i, \dots, G_j^i, \dots, G_{n_i}^i\}$, 树型结构引导稀疏的正则化表达形式如下:

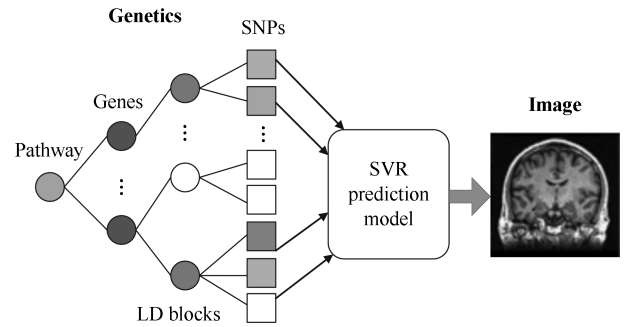


图 2 树型结构引导稀疏回归模型^[54]

Fig. 2 Tree-guided sparse regression model^[54]

$$\Omega_{\text{treeLasso}}(\mathbf{w}) = \sum_{i=1}^d \sum_{j=1}^{n_i} \alpha_j^i \|\mathbf{w}_{G_j^i}\|_2 \quad (6)$$

其中, α_j^i 为根据先验知识预先定义的任一结点 G_j^i 的权重, $\mathbf{w}_{G_j^i}$ 为最终需要优化学习出的树型结构中任一结点 G_j^i 的权重. 值得注意的是, 当一个结点的权重为零时, 其子节点也全部为零, 即该子树的全部特征与回归任务无关, 均没有被选择到. 相比传统的 Lasso 方法, 该模型优化得到的 SNP 特征在预测大脑灰质体积上具有较小的误差, 同时识别出这些与 MRI 脑区相关联的 SNP 位点具有层次结构的聚类

特性. 因此, 该模型可以扩展并应用到其他具有层次聚类特性的影像遗传学问题.

2.2 多变量影像回归单变量基因

在基于统计学习的影像遗传学研究中, 大部分的工作集中在发现和检测与影像相关联的多变量基因位点, 而很少有研究探索当表型测量变量变化时 SNP 值的变化情况, 即利用表型 QT 特征回归基因型 SNP 的值. Wang 等提出了任务相关的纵向稀疏回归模型来实现纵向影像表型与基因型的关联分析^[56]. 该回归模型损失函数的表达式如下:

$$L(W) = \|Y - X \otimes W\|_F^2 = \sum_{t=1}^T \|Y - X_t W_t\|_F^2 \quad (7)$$

其中, X 为影像数据, Y 为基因数据, 该模型需要学习优化的关联系数是一个张量形式 $W = \{W_1, \dots, W_T\} \in \mathbf{R}^{q \times p \times T}$, d 为影像表型变量特征数, p 为回归任务 SNP 的个数, T 为纵向时间点数 ($T = 4$, 分别为基线、6 个月、12 个月和 24 个月). 正则化约束表示形式如下:

$$\Omega_{\text{longitudinal}}(W) = \sum_{k=1}^q \sqrt{\sum_{t=1}^T \|\mathbf{w}_t^k\|_2^2} \quad (8)$$

实际上, $\Omega_{\text{longitudinal}}(W)$ 是 L_{21} 范数的一种推广形式, 如图 3 所示, 通过在多个回归任务以及多个时间点上特征权重的联合约束, 能够选择出与任务相关的纵向影像表型特征标志. 当 $p = 1$ 时, Y 表示风险基因 SNP 位点, 该模型可以从表型到基因型分析的新视角来研究单个基因对大脑结构和功能变化的影响.

在单变量基因多变量影像关联分析中, 绝大多数的研究都是针对单模态表型 QT. 而为研究基因与多模态脑影像之间的关联, Hao 等通过构建从脑影像 X 到基因 Y 的回归模型并引入 L_{21} 范数作为正则化约束的方法实现了多模态影像与候选风险基因位点的关联分析^[57], 其正则化表达式如下:

$$\Omega_{\text{multimodality}}(W) = \Omega_{L_{2,1}}(W) = \|W\|_{2,1} = \sum_i \sqrt{\sum_j \mathbf{w}_{ij}^2} \quad (9)$$

如图 4 所示, $W = [\mathbf{w}_{VBM}, \mathbf{w}_{FDG}, \mathbf{w}_{AV45}] \in \mathbf{R}^{q \times 3}$ 为多模态影像 (VBM、FDG、AV45) 与候选风险基因 APOE e4 的关联权重矩阵, 在此问题中 $j = \{1, 2, 3\}$, $i = \{1, \dots, q\}$. 该方法通过广义线性回归函数来实现与风险基因关联的多模态影像生物标记的特征选择. 这种基于多模态的影像关联分析

可以检测到鲁棒的一致性脑区, 相比单模态关联分析方法具有很强的抗噪声能力. 因此, 该方法可以应用到其他风险基因与多模态影像的关联分析中.

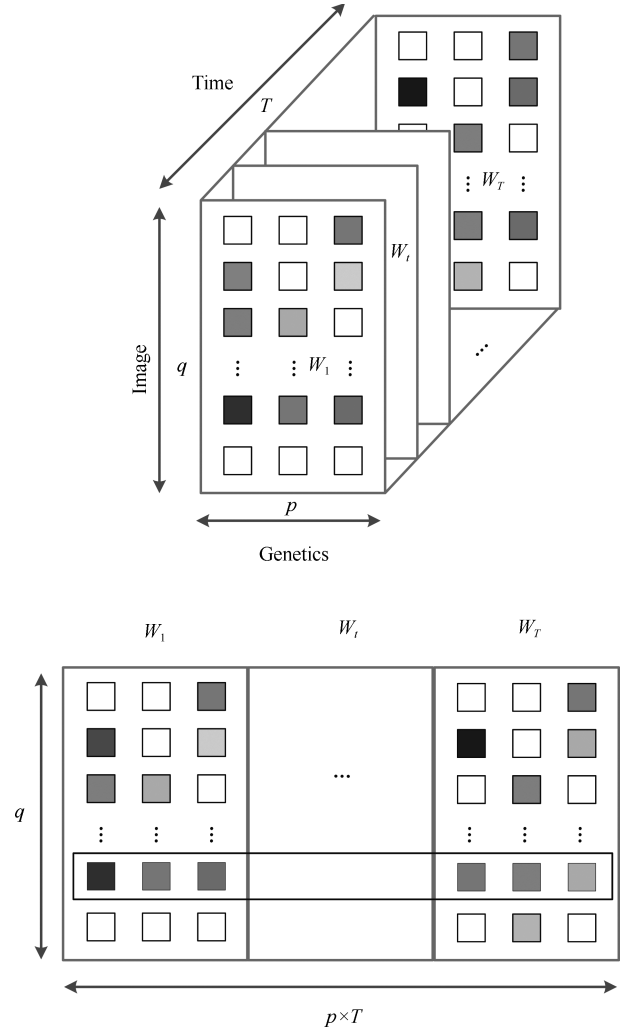
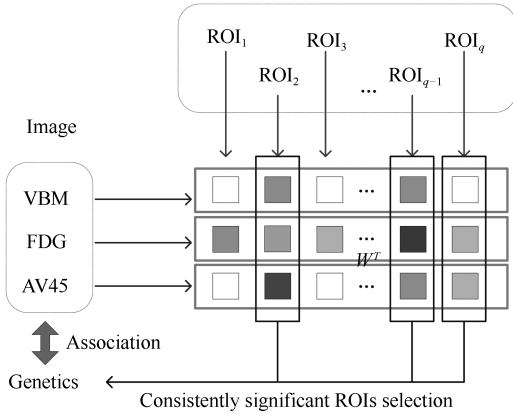
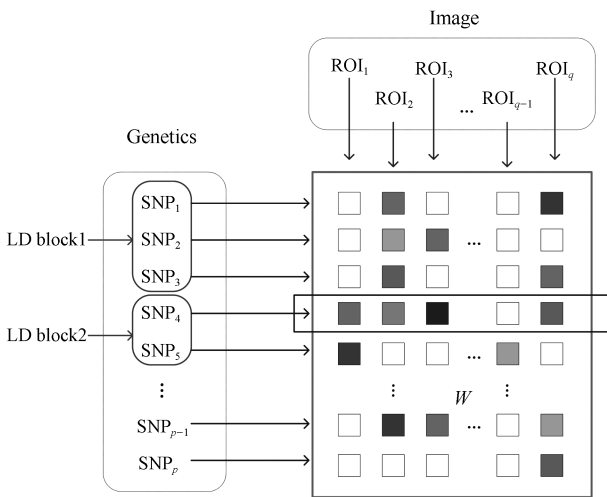


图 3 任务相关的纵向稀疏回归模型^[56]

Fig. 3 Task-correlated longitudinal sparse regression model^[56]

2.3 多变量基因回归多变量影像

上述的基因位点多变量分析只是针对于单个影像 QT 的回归, 并不能充分利用影像多变量特征之间的相关关系. 近年来, 也有一些方法是专门解决多变量基因输入多变量影像输出的问题, 如组稀疏多任务回归和特征选择模型^[58], 如图 5 所示. 多任务回归使用了本文第 2.2 节中多模态特征选择相同的表达式 L_{21} 范数 $\Omega_{L_{2,1}}(\mathbf{w})$ 作为一种可以用来约束多个联合相关表型与基因变量产生关联的正则化项, 表达式如下:

图 4 多模态关联模型^[57]Fig. 4 Multi-modality association model^[57]图 5 组稀疏多任务回归和特征选择模型^[58]Fig. 5 Group-sparse multi-task regression and feature selection model^[58]

$$\Omega_{\text{multitask}}(W) = \Omega_{L2,1}(W) = \|W\|_{2,1} = \sum_i \sqrt{\sum_j w_{ij}^2} \quad (10)$$

此外, Wang 等还考虑到 SNP 位点之间的连锁不平衡 LD 结构关系, 利用 $\|W\|_{G2,1}$ 正则化项在模型中嵌入 SNP 的成组关系这一先验信息, 使得在同一个 LD 组中的 SNP 被同时检测到, 表达形式如下:

$$\Omega_{G2,1}(W) = \|W\|_{G2,1} = \sum_{k=1}^K \sqrt{\sum_{i \in \pi_k} \sum_j w_{ij}^2} \quad (11)$$

实验结果表明这些嵌入多变量基因结构的稀疏学习模型所筛选出的位点对于回归任务具有较小的误差,

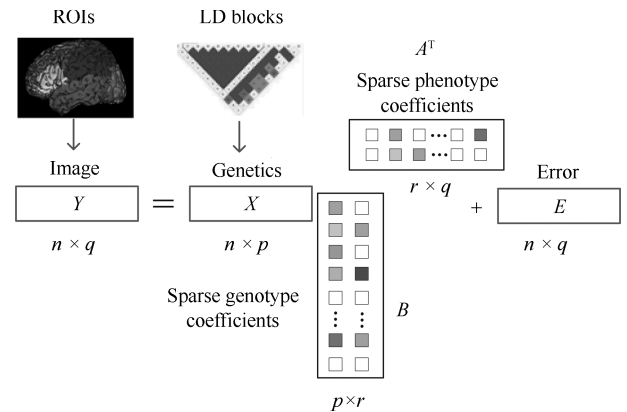
⁶<http://www.imperial.ac.uk/~gmontana/psrrr.htm>

从而在关联分析中具有更好的性能. 该模型能够检测到多个相关的基因位点与多个相关脑区的联合关联.

另外一种低秩回归 (Reduced rank regression, RRR) 的模型作为广义线性回归模型的扩展, 也可以解决多变量基因输入和多变量影像标记输出的关联分析问题. 其模型的表达形式如下:

$$\min_{A,B} \text{Tr}(Y - XBA^T)\Gamma(XBA^T)^T \quad (12)$$

其中, 权重矩阵 $W \in \mathbf{R}^{p \times q}$ 满足 $\text{rank}(W) \leq \min(p, q)$, 即权重矩阵可以分解成两个秩为 r 的满秩矩阵 $W = BA^T$, 其中 $B \in \mathbf{R}^{p \times r}$, $A \in \mathbf{R}^{q \times r}$. 权重矩阵 W 的分解不但可以减少关联分析中需要估计的参数, 而且可以分别对基因和影像变量进行稀疏化约束. 稀疏的低秩回归模型 (Sparse reduced rank regression, SRRR)⁶, 如图 6 所示, 通过使用 $L1$ 范数分别对 A 和 B 进行约束实现对相关 SNP 和影像 QT 的特征选择^[59-60]. 为方便求解, 可以假设 $X^T X = I$ 以及 $\Gamma = I$, SRRR 的目标函数形式如下:

图 6 稀疏低秩回归模型^[59-60]Fig. 6 Sparse reduced rank regression model^[59-60]

$$\min_{A,B} -2\text{Tr}A^T Y^T X B + \text{Tr}A^T A B^T B + \lambda_1 \|A\|_1 + \lambda_2 \|B\|_1 \quad (13)$$

在实际应用中常取 $r = 1$, 即 A 和 B 分别为影像和基因特征的权重向量. 实验结果表明, 相对多次单变量线性模型, SRRR 模型能够进行多变量基因型输入多变量表现型输出的关联分析, 在检测和识别相关变量问题中具有更好的性能. 该模型还可以通过引入更加丰富的生物学过程等先验信息进行扩展.

2.4 多变量基因多变量影像双变量关联分析

在影像遗传学研究中, 多变量回归模型已经能够很好地解决基因或者影像预测因子特征选择的问题; 而对于多输出的变量特征, 多任务回归模型^[58, 61]可以在一定程度上考虑多个回归输出变量之间的协方差结构关系. 但是高维的回归输出会产生较高的计算时间代价, 而且多变量输出结构复杂, 仅简单地考虑对多个回归任务进行成组约束进行关联分析的模型往往过于严格和理想化. 为了充分考虑双多变量之间的协方差结构, Liu 等提出了使用并行独立成分分析 (Parallel independent component analysis, PICA)^[62–63]方法来分析具有关联机制的基因与影像数据, 从而发现这两种模态数据的最大相关独立成分. 但是该方法并不能恢复产生贡献的 SNP 和影像中的重要脑区, 从而导致这些成分丧失了合理的生物标记解释性. 另外一种类型的双多变量模型, 如典型相关分析 (Canonical correlation analysis, CCA)^[64–65]或者偏最小二乘回归 (Partial least squares regression, PLS)^[66–67]可以分别寻找两组变量中的线性组合使得基因和影像两块数据之间的相关性或者协方差最大, 相对回归模型, 可以更好地解决多变量基因多变量影像关联分析这一问题. 在此以经典的 CCA 模型为例阐述双多变量关联方法, 该类模型能够找到两组模态数据特征变量之间的相关性, 其表达形式如下:

$$\begin{aligned} & \max_{\mathbf{u}, \mathbf{v}} \mathbf{u}^T \mathbf{X}^T \mathbf{Y} \mathbf{v} \\ & \text{s.t. } \mathbf{u}^T \mathbf{X}^T \mathbf{X} \mathbf{u} = 1, \mathbf{v}^T \mathbf{Y}^T \mathbf{Y} \mathbf{v} = 1 \end{aligned} \quad (14)$$

其中, \mathbf{u} 和 \mathbf{v} 分别为 SNP 和 QT 两组模态数据特征的权重向量. 但是在高维数据中, 特征变量往往存在噪声和冗余, 即并不是所有的 SNP 和 QT 特征变量都存在关联作用. 因此, 为筛选出具有解释意义的少量相关基因和影像特征, 很多工作也将结构稀疏化特征选择的思路引入了经典的双变量关联分析中, 即稀疏典型相关分析 (Sparse canonical correlation analysis, SCCA)^[68–70]和稀疏偏最小二乘回归 (Sparse partial least squares regression, SPLS)^[71–72]. 在此, 以 SCCA 模型为例进行阐述, 该类方法在 CCA 的基础上利用了 $L1$ 范数分别对双多变量权重向量进行稀疏性约束, 表达形式如下:

$$\Omega_{\text{Lasso}}(\mathbf{u}) = \|\mathbf{u}\|_1 = \sum_i |\mathbf{u}_i| \leq c_1 \quad (15)$$

$$\Omega_{\text{Lasso}}(\mathbf{v}) = \|\mathbf{v}\|_1 = \sum_j |\mathbf{v}_j| \leq c_2 \quad (16)$$

其中 c_1 和 c_2 分别为控制特征向量 \mathbf{u} 和 \mathbf{v} 稀疏性的

参数. 将稀疏学习引入多基因多影像关联分析中, 模型可以自动地从高维的双变量中选择具有相关性的稀疏 SNP 和 QT 特征变量. 然而, SCCA 的一个主要问题是该模型仍然没有充分考虑特征变量之间的结构关系, 即很多先验信息并没有被用到模型的建立中, 例如同一个 LD 块中基因位点 SNP 之间可能具有某些共同的特性, 以及大脑在完成某种功能时需要多个脑区协同工作. 因此, 在多基因多影像关联的研究中, 为了弥补传统 SCCA 的不足, 很多学者利用各种先验信息对 SCCA 模型进行了扩展和改进^[73–76]. 例如, Yan 等提出了嵌入淀粉样蛋白转录先验信息的结构化 SCCA 模型⁸, 如图 7 所示. 该模型利用组稀疏正则化约束将 SNP 位点的 LD 块成组的先验信息嵌入 SCCA 优化模型中^[73], 表达形式如下:

$$\Omega_{\text{groupLasso}}(\mathbf{u}) = \sum_{i=1}^g \sqrt{\sum_{j \in G(i)} \mathbf{u}_j^2} \leq c_3 \quad (17)$$

其中, $L2$ 范数约束组内的特征变量尽可能具有相同的权重贡献, 即在关联分析中更倾向于选择同一个 LD 块中的 SNP 位点; $L1$ 范数通过约束组间的稀疏性来选择少数具有强关联的 LD 块. 此外, 该模型还引入了大脑功能网络信息作为脑区变量特征相似性的先验知识, 即当大脑网络中的连接权重较高时则两个脑区结点具有相似的特性 (基因表达高度相关), 其正则化约束表达形式如下:

$$\begin{aligned} \Omega_{\text{network}}(\mathbf{v}) &= \sum_{(i,j) \in E, i < j} \tau(\mathbf{w}_{ij}) \times \\ & \|\mathbf{v}_i - \text{sign}(\mathbf{w}_{ij}) \mathbf{v}_j\|_2 \leq c_4 \end{aligned} \quad (18)$$

其中 \mathbf{v}_i 与 \mathbf{v}_j 分别表示大脑网络上的任意两个结点特征权重. $\text{sign}(\mathbf{w}_{ij})$ 为 \mathbf{v}_i 与 \mathbf{v}_j 相关性的符号, $\text{sign}(\mathbf{w}_{ij})$ 为正时, \mathbf{v}_i 与 \mathbf{v}_j 的特征被拉近呈正相关关系; $\text{sign}(\mathbf{w}_{ij})$ 为负时, \mathbf{v}_i 与 \mathbf{v}_j 的特征差异相反呈负相关关系. $\tau(\mathbf{w}_{ij})$ 为 \mathbf{v}_i 与 \mathbf{v}_j 的连接强度, $\tau(\mathbf{w}_{ij})$ 的强度越高表示了 \mathbf{v}_i 与 \mathbf{v}_j 倾向于这两个脑区特征变量同时被选择. 值得注意的是, Ω_{network} 是上文提到的 $\Omega_{\text{fusedLasso}}$ 的一种扩展形式. 实验结果表明, 相比传统的 SCCA, 结构化的扩展模型通过嵌入 SNP 位点 LD 成组信息以及淀粉样蛋白引导的大脑结构特性所选择的对应特征具有更强的基因影像关联性和生物解释意义.

⁷<http://mialab.mrn.org/software/fit/index.html>

⁸<http://www.iu.edu/~hdbig/SCCA/>

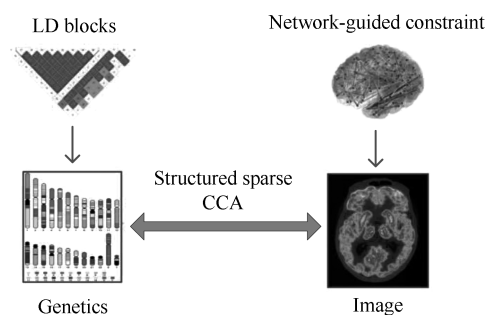


图 7 稀结构化稀疏的双多变量关联模型^[73]

Fig. 7 Structured sparse bi-multivariate correlation model^[73]

3 总结和展望

影像遗传学作为新兴的前沿交叉学科, 涉及到了神经科学、影像学、遗传学、医学、生物统计学、数据挖掘、统计学习等多种科学和研究技术. 而基因组、多模态影像组数据 (包括不同时间点的纵向脑影像) 也为影像遗传学研究提供了丰富的数据实验平台, 使得基因与脑结构或者功能之间产生关联的致病机制能够通过具有遗传特性的影像内表型呈现出来. 统计学习技术作为基于数据驱动的关联分析强有力工具, 通过充分发掘和利用基因与影像等生物标记数据内在的结构信息, 能够分析易感基因与大脑结构或者功能的相关性, 更好地揭示脑认知行为或者相关疾病的产生机制. 本文回顾了近年来基于稀疏学习的关联分析算法在影像遗传学研究领域的应用, 大量的实验和报告显示模型所检测到的部分关联结果也在生物和医学界得到了验证.

本文所回顾的基于结构化方法的多变量影像遗传学研究之所以可以取得很好的效果, 是因为数据分析的模型中嵌入了大量的先验知识 (例如 LD 能够刻画 SNP 之间的简单结构关系). 而本文所归纳总结的结构化多变量方法正是通过生物学过程以及医学领域知识 (如代谢通路/网络、多模态融合、诊断信息等) 诱导的一般性方法. 在此基础上, 研究者可以对先验信息进行补充、对模型进行拓展. 目前有一些工作已经考虑了利用更具有遗传功能生物特性的先验知识应用在模型的建立和学习训练中, 其中包括基因本体 (Gene ontology, GO)、功能标注、通路分析系统 (如 KEGG (Kyoto encyclopedia of genes and genomes) 通路数据库或者 OMIM (Online mendelian inheritance in man) 疾病数据库等)^[52, 77]. 那么如何根据这些先验知识设计更加适合实际应用问题的模型进行数据分析, 即实现假设驱动与数据驱动相结合^[78], 以期待得到更好的关联结果依然是当前的研究热点.

尽管结构化多基因多影像性状的关联结果可以在一定程度上解释遗传效应, 但是对于影响同一性状表现的多个非等位基因之间可能存在交互关系即异位显性 (Epistasis) 的机制并不是十分清楚. 目前已经有一些工作是针对 SNP 之间的交互作用对影像 QT 的研究^[79]. 这些方法主要是基于遍历成对搜索的方法. 例如, Hibar 等利用迭代确定独立筛选 (Sure independence screening, SIS) 算法实现并检测出与某个脑区性状显著关联的 SNP-SNP 之间的交互作用^[80]. 这些遍历搜索带来相当高的计算时间代价, 而一些高效的稀疏模型^[49] 对于多次交互关系的异位显性研究有望提供高效的学习算法. 对于高维特征变量基因位点检测的效率问题, 除了提高算法本身的效率以外, 还可以通过引入分布式并行计算方法完成大数据的计算^[81]. 因此, 还需要进一步发展和构建更加高效的算法模型或者工作框架来进行全基因组和全脑特征变量交互的影像遗传学研究.

在以往的影像遗传学研究中, 大多数工作采用无监督的关联分析, 即仅仅关注基因与脑结构或者功能之间的关系, 忽略了诊断类别或相关的量表得分信息. 而目前有一些研究工作已经通过引入诊断信息来指导模型进行基因-影像关联分析. 例如, Vounou 等采用“两步”策略, 首先选择具有判别性的脑影像区域, 然后再进行其与基因的关联分析^[60]. 随后, Batmanghelich 为了避免“两步”策略的关联信息损失, 基于贝叶斯框架设计了融合基因、影像、诊断三者逻辑联系的模型, 将脑影像特征作为一个中间表型来检测与疾病相关的遗传变异, 同时地进行基因和影像特征的选择^[82]; 这种产生式建模方法, 带来了良好的关联解释性, 但是在具体实现中计算复杂程度较高. Hao 等通过构建基于多类别诊断标记信息引导的多模态结构和功能影像的模型来检测与候选风险基因位点相关联的一致性脑区^[83]; 但是对于这样的单输出回归模型, 并不能用于发现和检测与影像相关的多变量基因位点. 上述基于引入监督信息的基因与影像关联的研究目的是为了获得脑认知行为或者疾病产生的机制, 为疾病的诊断和预测提供依据, 但是最终并没有显式地应用在疾病的诊断预测中; 另一方面, 也有一些研究只关注疾病诊断分类问题, 即直接将基因与影像等多模态数据进行简单的特征融合作为预测因子以提高分类精度^[42, 61, 84], 但其中并没有涉及多模态数据之间的相关性分析而缺乏复杂致病机制的解释性. 因此, 如何对基因、脑影像、量表得分、诊断信息数据构建联合关联、回归和分类的多任务统一模型^[85], 既能揭示基因与脑影像之间的关联性同时又实现基于生物标志特征的疾病诊断和预测, 也将成为影像遗传学未来研究的发展方向.

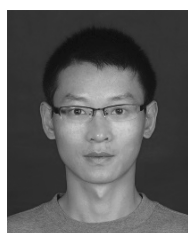
References

- Hariri A R, Weinberger D R. Imaging genomics. *British Medical Bulletin*, 2003, **65**(1): 259–270
- Thompson P M, Martin N G, Wright M J. Imaging genomics. *Current Opinion in Neurology*, 2010, **23**(4): 368–373
- Glahn D C, Thompson P M, Blangero J. Neuroimaging endophenotypes: strategies for finding genes influencing brain structure and function. *Human Brain Mapping*, 2007, **28**(6): 488–501
- Gottesman I I, Gould T D. The endophenotype concept in psychiatry: etymology and strategic intentions. *The American Journal of Psychiatry*, 2003, **160**(4): 636–645
- Meyer-Lindenberg A, Weinberger D R. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nature Reviews Neuroscience*, 2006, **7**(10): 818–827
- Ge T, Schumann G, Feng J F. Imaging genetics — towards discovery neuroscience. *Quantitative Biology*, 2013, **1**(4): 227–245
- Winkler A M, Kochunov P, Blangero J, Almasy L, Zilles K, Fox P T, Duggirala R, Glahn D C. Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *NeuroImage*, 2010, **53**(3): 1135–1146
- Smith S M, Fox P T, Miller K L, Glahn D C, Fox P M, Mackay C E, Filippini N, Watkins K E, Toro R, Laird A R, Beckmann C F. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the United States of America*, 2009, **106**(31): 13040–13045
- Tost H, Bilek E, Meyer-Lindenberg A. Brain connectivity in psychiatric imaging genetics. *NeuroImage*, 2012, **62**(4): 2250–2260
- Rubinov M, Sporns O. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage*, 2010, **52**(3): 1059–1069
- Hardy J, Singleton A. Genomewide association studies and human disease. *The New England Journal of Medicine*, 2009, **360**(17): 1759–1768
- Klein R J, Zeiss C, Chew E Y, Tsai J Y, Sackler R S, Haynes C, Henning A K, SanGiovanni J P, Mane S M, Mayne S T, Bracken M B, Ferris F L, Ott J, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. *Science*, 2005, **308**(5720): 385–389
- Esslinger C, Walter H, Kirsch P, Erk S, Schnell K, Arnold C, Haddad L, Mier D, von Boerfeld C O, Raab K, Witt S H, Rietschel M, Cichon S, Meyer-Lindenberg A. Neural mechanisms of a genome-wide supported psychosis variant. *Science*, 2009, **324**(5927): 605
- Medland S E, Jahanshad N, Neale B M, Thompson P M. Whole-genome analyses of whole-brain data: working within an expanded search space. *Nature Neuroscience*, 2014, **17**(6): 791–800
- Liu J Y, Calhoun V D. A review of multivariate analyses in imaging genetics. *Frontiers in Neuroinformatics*, 2014, **8**: Article No. 29
- Thompson P M, Ge T, Glahn D C, Jahanshad N, Nichols T E. Genetics of the connectome. *NeuroImage*, 2013, **80**: 475–488
- Daniel W W, Cross C L. *Biostatistics: A Foundation for Analysis in the Health Sciences* (Tenth Edition). New York: Wiley, 2013.
- Potkin S G, Guffanti G, Lakatos A, Turner J A, Kruggel F, Fallon J H, Saykin A J, Orro A, Lupoli S, Salvi E, Weiner M, Macciardi F, The Alzheimer's Disease Neuroimaging Initiative. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS One*, 2009, **4**(8): Article No. e6501
- Shen L, Thompson P M, Potkin S G, Bertram L, Farrer L A, Foroud T M, Green R C, Hu X L, Huentelman M J, Kim S, Kauwe J S K, Li Q Q, Liu E C, Macciardi F, Moore J H, Munsie L, Nho K, Ramanan V K, Risacher S L, Stone D J, Swaminathan S, Toga A W, Weiner M W, Saykin A J. Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. *Brain Imaging and Behavior*, 2014, **8**(2): 183–207
- Risacher S L, Shen L, West J D, Kim S, McDonald B C, Beckett L A, Harvey D J, Jack Jr C R, Weiner M W, Saykin A J. Longitudinal MRI atrophy biomarkers: relationship to conversion in the ADNI cohort. *Neurobiology of Aging*, 2010, **31**(8): 1401–1418
- Risacher S L, Kim S, Shen L, Nho K, Foroud T, Green R C, Petersen R C, Jack Jr C R, Aisen P S, Koeppe R A, Jagust W J, Shaw L M, Trojanowski J Q, Weiner M W, Saykin A J. The role of apolipoprotein E (APOE) genotype in early mild cognitive impairment (E-MCI). *Frontiers in Aging Neuroscience*, 2013, **5**: Article No. 11
- Ho A J, Stein J L, Hua X, Lee S, Hibar D P, Leow A D, Dinov I D, Toga A W, Saykin A J, Shen L, Foroud T, Pankratz N, Huentelman M J, Craig D W, Gerber J D, Allen A N, Corneveaux J J, Stephan D A, DeCarlis C S, DeChairo B M, Potkin S G, Jack Jr C R, Weiner M W, Raji C A, Lopez O L, Becker J T, Carmichael O T, Thompson P M. A commonly carried allele of the obesity-related FTO gene is associated with reduced brain volume in the healthy elderly. *Proceedings of the National Academy of Sciences of the United States of America*, 2010, **107**(18): 8404–8409
- Reiman E M, Chen K W, Liu X F, Bandy D, Yu M X, Lee D, Ayutyanont N, Keppler J, Reeder S A, Langbaum J B S, Alexander G E, Klunk W E, Mathis C A, Price J C, Aizenstein H J, DeKosky S T, Caselli R J. Fibrillar amyloid- β burden in cognitively normal people at 3 levels of genetic risk for Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America*, 2009, **106**(16): 6820–6825
- Sloan C D, Shen L, West J D, Wishart H A, Flashman L A, Rabin L A, Santulli R B, Guerin S J, Rhodes C H, Tsongalis G J, McAllister T W, Ahles T A, Lee S L, Moore J H, Saykin A J. Genetic pathway-based hierarchical clustering analysis of older adults with cognitive complaints and amnesic mild cognitive impairment using clinical and neuroimaging phenotypes. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*, 2010, **153B**(5): 1060–1069
- Swaminathan S, Shen L, Risacher S L, Yoder K K, West J D, Kim S, Nho K, Foroud T, Inlow M, Potkin S G, Huentelman M J, Craig D W, Jagust W J, Koeppe R A, Mathis C A, Jack Jr C R, Weiner M W, Saykin A J. Amyloid pathway-based candidate gene analysis of [^{11}C]PiB-PET in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. *Brain Imaging and Behavior*, 2012, **6**(1): 1–15
- Chiang M C, Barysheva M, McMahon K L, de Zubizaray G I, Johnson K, Montgomery G W, Martin N G, Toga A W, Wright M J, Shapshak P, Thompson P M. Gene network effects on brain microstructure and intellectual performance identified in 472 twins. *Journal of Neuroscience*, 2012, **32**(25): 8732–8745

- 27 Saykin A J, Shen L, Foroud T M, Potkin S G, Swaminathan S, Kim S, Risacher S L, Nho K, Huentelman M J, Craig D W, Thompson P M, Stein J L, Moore J H, Farrer L A, Green R C, Bertram L, Jack Jr C R, Weiner M W. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans. *Alzheimer's & Dementia*, 2010, **6**(3): 265–273
- 28 Potkin S G, Turner J A, Fallon J A, Lakatos A, Keator D B, Guffanti G, Macciardi F. Gene discovery through imaging genetics: identification of two novel genes associated with schizophrenia. *Molecular Psychiatry*, 2009, **14**(4): 416–428
- 29 Shen L, Kim S, Risacher S L, Nho K, Swaminathan S, West J D, Foroud T, Pankratz N, Moore J H, Sloan C D, Huentelman M J, Craig D W, DeChairo B M, Potkin S G, Jack Jr C R, Weiner M W, Saykin A J. 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
- 30 Stein J L, Hua X, Lee S, Ho A J, Leow A D, Toga A W, Saykin A J, Shen L, Foroud T, Pankratz N, Huentelman M J, Craig D W, Gerber J D, Allen A N, Corneveaux J J, DeChairo B M, Potkin S G, Weiner M W, Thompson P M. Voxelwise genome-wide association study (vGWAS). *NeuroImage*, 2010, **53**(3): 1160–1174
- 31 Biffi A, Anderson C D, Desikan R S, Sabuncu M, Cortellini L, Schmansky N, Salat D, Rosand J, Alzheimer's Disease Neuroimaging Initiative (ADNI). Genetic variation and neuroimaging measures in Alzheimer disease. *Archives of Neurology*, 2010, **67**(6): 677–685
- 32 Kauwe J S K, Bertelsen S, Mayo K, Cruchaga C, Abraham R, Hollingworth P, Harold D, Owen M J, Williams J, Lovestone S, Morris J C, Goate A M. Suggestive synergy between genetic variants in TF and HFE as risk factors for Alzheimer's disease. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*, 2010, **153B**(4): 955–959
- 33 Dickerson B C, Wolk D A. Dysexecutive versus amnesic phenotypes of very mild Alzheimer's disease are associated with distinct clinical, genetic and cortical thinning characteristics. *Journal of Neurology, Neurosurgery & Psychiatry*, 2011, **82**(1): 45–51
- 34 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M A R, Bender D, Maller J, Sklar P, de Bakker P I W, Daly M J, Sham P C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 2007, **81**(3): 559–575
- 35 Gombor S, Jung H J, Dong F, Calder B, Atzmon G, Barzilai N, Tian X L, Pothof J, Hoeijmakers J H J, Campisi J, Vijg J, Suh Y. Comprehensive microRNA profiling in B-cells of human centenarians by massively parallel sequencing. *BMC Genomics*, 2012, **13**(1): Article No. 353
- 36 Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 2001, **29**(4): 1165–1188
- 37 Hibar D P, Stein J L, Kohannim O, Jahanshad N, Saykin A J, Shen L, Kim S, Pankratz N, Foroud T, Huentelman M J, Potkin S G, Jack Jr C R, Weiner M W, Toga A W, Thompson P M. Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. *NeuroImage*, 2011, **56**(4): 1875–1891
- 38 Hibar D P, Stein J L, Kohannim O, Jahanshad N, Jack C R, Weiner M W, Toga A W, Thompson P M. Principal components regression: multivariate, gene-based tests in imaging genomics. In: Proceedings of the 2011 IEEE International Symposium on Biomedical Imaging: from Nano to Macro. Chicago, IL, USA: IEEE, 2011. 289–293
- 39 Hibar D P, Kohannim O, Stein J L, Chiang M C, Thompson P M. Multilocus genetic analysis of brain images. *Frontiers in Genetics*, 2011, **2**: Article No. 73
- 40 Ye J P, Liu J. Sparse Methods for Biomedical Data. *ACM Sigkdd Explorations Newsletter*, 2012, **14**(1): 4–15
- 41 Wang J, Yang T, Thompson P, Ye J. Sparse models for imaging genetics. *Machine Learning and Medical Imaging*. New York: Academic Press, 2016. 129–151
- 42 Lin D D, Cao H B, Calhoun V D, Wang Y P. Sparse models for correlative and integrative analysis of imaging and genetic data. *Journal of Neuroscience Methods*, 2014, **237**: 69–78
- 43 Yan J, Du L, Yao X, Shen L. Machine learning in brain imaging genomics. *Machine Learning and Medical Imaging*. New York: Academic Press, 2016. 411–434
- 44 Donoho D L. Compressed sensing. *IEEE Transactions on Information Theory*, 2006, **52**(4): 1289–1306
- 45 Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, **58**(1): 267–288
- 46 Kohannim O, Hibar D P, Stein J L, Jahanshad N, Jack C R, Weiner M W, Toga A W, Thompson P M. Boosting power to detect genetic associations in imaging using multi-locus, genome-wide scans and ridge regression. In: Proceedings of the 2011 IEEE International Symposium on Biomedical Imaging: from Nano to Macro. Chicago, IL, USA: IEEE, 2011. 1855–1859
- 47 Kohannim O, Hibar D P, Jahanshad N, Stein J L, Hua X, Toga A W, Jack C R, Weiner M W, Thompson P M. Predicting temporal lobe volume on MRI from Genotypes Using L^1 - L^2 regularized regression. In: Proceedings of the 9th IEEE International Symposium on Biomedical Imaging (ISBI). Barcelona, Spain: IEEE, 2012. 1160–1163
- 48 Kohannim O, Hibar D P, Stein J L, Jahanshad N, Hua X, Rajagopalan P, Toga A W, Jack Jr C R, Weiner M W, de Zubicaray G I, McMahon K L, Hansell N K, Martin N G, Wright M J, Thompson P M, The Alzheimer's Disease Neuroimaging Initiative. Discovery and replication of gene influences on brain structure using LASSO regression. *Frontiers in Neuroscience*, 2012, **6**: Article No. 115
- 49 Yang T, Wang J, Sun Q, Hibar D P, Jahanshad N, Liu L, Wang Y L, Zhan L, Thompson P M, Ye J P. Detecting genetic risk factors for Alzheimer's disease in whole genome sequence data via lasso screening. In: Proceedings of the 12th International Symposium on Biomedical Imaging (ISBI). New York, USA: IEEE, 2015. 985–989
- 50 Silver M, Montana G. Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. *Statistical Applications in Genetics and Molecular Biology*, 2012, **11**(1): Article No. 7
- 51 Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B-Statistical Methodology*, 2006, **68**(1): 49–67
- 52 Silver M, Chen P, Li R Y, Cheng C Y, Wong T Y, Tai E S, Teo Y Y. Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two Asian cohorts. *PLoS Genetics*, 2013, **9**(11): Article No. e1003939
- 53 Barrett J C, Fry B, Maller J, Daly M J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 2005, **21**(2): 263–265

- 54 Hao X K, Yu J T, Zhang D Q. Identifying genetic associations with MRI-derived measures via tree-guided sparse learning. In: Proceedings of the 17th International Conference on Medical Image Computing and Computer-Assisted Intervention. Boston, USA: Springer, 2014. 757–764
- 55 Wang J, Ye J P. Multi-layer feature reduction for tree structured group lasso via hierarchical projection. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montréal, Quebec, Canada: MIT Press, 2015. 1279–1287
- 56 Wang H, Nie F P, Huang H, Yan J W, Kim S, Nho K, Risacher S L, Saykin A J, Shen L. From phenotype to genotype: an association study of longitudinal phenotypic markers to Alzheimer's disease relevant SNPs. *Bioinformatics*, 2012, **28**(18): i619–i625
- 57 Hao X K, Yan J W, Yao X H, Risacher S L, Saykin A J, Zhang D Q, Shen L. Diagnosis-guided method for identifying multi-modality neuroimaging biomarkers associated with genetic risk factors in Alzheimer's disease. In: Proceedings of the Pacific Symposium on Biocomputing. Kohala Coast, Hawaii, USA: Stanford, 2016. 108–119
- 58 Wang H, Nie F P, Huang H, Kim S, Nho K, Risacher S L, Saykin A J, Shen L, The Alzheimer's Disease Neuroimaging Initiative. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics*, 2012, **28**(2): 229–237
- 59 Vounou M, Nichols T E, Montana G, The Alzheimer's Disease Neuroimaging Initiative. Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. *NeuroImage*, 2010, **53**(3): 1147–1159
- 60 Vounou M, Janouseva E, Wolz R, Stein J L, Thompson P M, Rueckert D, Montana G, The Alzheimer's Disease Neuroimaging Initiative. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *NeuroImage*, 2012, **60**(1): 700–716
- 61 Wang H, Nie F P, Huang H, Risacher S L, Saykin A J, Shen L, The Alzheimer's Disease Neuroimaging Initiative. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 2012, **28**(1): i127–i136
- 62 Liu J Y, Pearlson G, Windemuth A, Ruano G, Perrone-Bizzozero N I, Calhoun V. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Human Brain Mapping*, 2009, **30**(1): 241–255
- 63 Meda S A, Narayanan B, Liu J Y, Perrone-Bizzozero N I, Stevens M C, Calhoun V D, Glahn V D, Shen L, Risacher S L, Saykin A J, Pearlson G D. A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer's disease in the ADNI cohort. *NeuroImage*, 2012, **60**(3): 1608–1621
- 64 Hotelling H. The most predictable criterion. *Journal of Educational Psychology*, 1935, **26**(2): 139–142
- 65 Correa N M, Li Y O, Adali T, Calhoun V D. Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia. *IEEE Journal of Selected Topics in Signal Processing*, 2008, **2**(6): 998–1007
- 66 Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method. *Matrix Pencils*. Berlin, Heidelberg: Springer, 1983: 286–293
- 67 Krishnan A, Williams L J, McIntosh A R, Abdi H. Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. *NeuroImage*, 2011, **56**(2): 455–475
- 68 Witten D M, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 2009, **10**(3): 515–534
- 69 Lê Cao K A, Martin P G P, Robert-Granié C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 2009, **10**(1): Article No. 34
- 70 Chi E C, Allen G I, Zhou H, Kohannim O, Lange K, Thompson P M. Imaging genetics via sparse canonical correlation analysis. In: Proceedings of the 10th International Symposium on Biomedical Imaging (ISBI). San Francisco, CA, USA: IEEE, 2013. 740–743
- 71 Le Floch É, Guillemot V, Frouin V, Pinel P, Lalanne C, Trinchera L, Tenenhaus A, Moreno A, Zilbovicius M, Bourgeron T, Dehaene S, Thirion B, Poline J B, Duchesnay É. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *NeuroImage*, 2012, **63**(1): 11–24
- 72 Lê Cao K A, Rossouw D, Robert-Granié C, Philippe B. A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 2008, **7**(1): 1–32
- 73 Yan J W, Du L, Kim S, Risacher S L, Huang H, Moore J H, Saykin A J, Shen L, The Alzheimer's Disease Neuroimaging Initiative. Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. *Bioinformatics*, 2014, **30**(17): i564–i571
- 74 Du L, Huang H, Yan J W, Kim S, Risacher S L, Inlow M, Moore J H, Saykin A J, Shen L, The Alzheimer's Disease Neuroimaging Initiative. Structured sparse canonical correlation analysis for brain imaging genetics: an improved Graphnet method. *Bioinformatics*, 2016, **32**(10): 1544–1551
- 75 Lin D D, Calhoun V D, Wang Y P. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. *Medical Image Analysis*, 2014, **18**(6): 891–902
- 76 Fang J, Lin D D, Schulz S C, Xu Z B, Calhoun V D, Wang Y P. Joint sparse canonical correlation analysis for detecting differential imaging genetics modules. *Bioinformatics*, 2016, **32**(22): 3480–3488
- 77 Yao X H, Yan J W, Kim S, Nho K, Risacher S L, Inlow M, Moore J H, Saykin A J, Shen L, The Alzheimer's Disease Neuroimaging Initiative. Two-dimensional enrichment analysis for mining high-level imaging genetic associations. *Brain Informatics and Health*. Cham: Springer, 2015. 115–124
- 78 Huys Q J M, Maia T V, Frank M J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 2016, **19**(3): 404–413
- 79 Birnbaum R, Weinberger D R. Functional neuroimaging and schizophrenia: a view towards effective connectivity modeling and polygenic risk. *Dialogues in Clinical Neuroscience*, 2013, **15**(3): 279–289
- 80 Hibar D P, Stein J L, Jahanshad N, Kohannim O, Toga A W, McMahon K L, de Zubicaray G I, Montgomery G W, Martin N G, Wright M J, Weiner M W, Thompson P M. Exhaustive search of the SNP-SNP interactome identifies epistatic effects on brain volume in two cohorts. In: Proceedings of the 16th International Conference on Medical Image Computing and Computer-Assisted Intervention. Nagoya, Japan: Springer, 2013. 600–607

- 81 Wang Y, Goh W, Wong L, Montana G, The Alzheimer's Disease Neuroimaging Initiative. Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes. *BMC Bioinformatics*, 2013, **14**(S16): Article No. S6
- 82 Batmanghelich N K, Dalca A V, Sabuncu M R, Golland P. Joint modeling of imaging and genetics. In: *Proceedings of the 23rd International Conference on Information Processing in Medical Imaging*. Asilomar, CA, USA: Springer, 2013. 766–777
- 83 Hao X K, Yao X H, Yan J W, Risacher S L, Saykin A J, Zhang D Q, Shen L. Identifying multimodal intermediate phenotypes between genetic risk factors and disease status in Alzheimer's disease. *Neuroinformatics*, 2016, **14**(4): 439–452
- 84 Cao H B, Duan J B, Lin D D, Calhoun V, Wang Y P. Integrating fMRI and SNP data for biomarker identification for schizophrenia with a sparse representation based variable selection method. *BMC Medical Genomics*, 2013, **6**(S3): Article No. S2
- 85 Gross S M, Tibshirani R. Collaborative regression. *Biostatistics*, 2015, **16**(2): 326–338



郝小可 河北工业大学计算机科学与软件学院讲师。于 2017 年在南京航空航天大学计算机科学与技术学院获得博士学位。分别于 2009 年和 2012 年在南京信息工程大学计算机与软件学院获得学士学位和硕士学位。主要研究方向为机器学习, 影像遗传学。

E-mail: robinhc@163.com

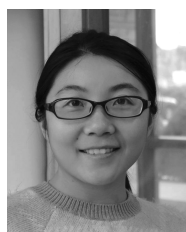
(HAO Xiao-Ke) Lecturer at the School of Computer Science and Engineering, Hebei University of Technology. He received his Ph. D. degree from the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics in 2017. He received his bachelor degree and master degree from the School of Computer and Software, Nanjing University of Information Science and Technology in 2009 and 2012, respectively. His research interest covers machine learning and imaging genetics.)



李蝉秀 南京航空航天大学计算机科学与技术学院硕士研究生。2015 年在南京航空航天大学计算机科学与技术学院获得学士学位。主要研究方向为机器学习, 影像遗传学。

E-mail: lcx_show@nuaa.edu.cn

(LI Chan-Xiu) Master student at the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. She received her bachelor degree from the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics in 2015. Her research interest covers machine learning and imaging genetics.)



严景文 印第安纳大学普渡大学印第安纳波利斯联合分校信息学与计算学院生物健康信息学系助理教授。曾分别在南京航空航天大学和华南理工大学获得学士学位和硕士学位。2015 年获得印第安纳大学信息学与计算学院博士学位。主要研究方向为机器学习, 影像遗传学。

E-mail: jingyan@iupui.edu

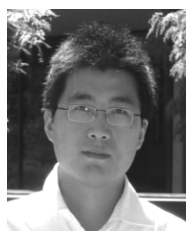
(YAN Jing-Wen) Assistant professor in the Department of BioHealth Informatics, School of Informatics and Computing, Indiana University Purdue University Indianapolis, USA. She received her bachelor degree from Nanjing University of Aeronautics and Astronautics. She received her master degree from Huazhong University of Science and Technology. She received her Ph. D. degree from the School of Informatics and Computing, Indiana University. Her research interest covers machine learning and imaging genetics.)



沈理 印第安纳大学医学院放射学与影像科学系副教授。曾分别在西安交通大学和上海交通大学获得学士和硕士学位, 在达特茅斯学院获得博士学位, 专业均为计算机科学。主要研究方向为医学影像计算, 信息生物学, 影像遗传学, 脑连接组。

E-mail: shenli@iu.edu

(SHEN Li) Associate professor of Radiology and Imaging Sciences at Indiana University School of Medicine. He received his bachelor degree from Xi'an Jiao Tong University, master degree from Shanghai Jiao Tong University, and Ph. D. degree from Dartmouth College, all in computer science. His research interest covers medical image computing, bioinformatics, imaging genomics, and brain connectomics.)



张道强 南京航空航天大学计算机科学与技术学院教授。分别于 1999 年和 2004 年在南京航空航天大学获得学士学位和博士学位。主要研究方向为机器学习, 模式识别, 数据挖掘以及医学影像分析。本文通信作者。

E-mail: dqzhang@nuaa.edu.cn

(ZHANG Dao-Qiang) Professor at the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. He received his bachelor degree and Ph. D. degree in computer science from Nanjing University of Aeronautics and Astronautics, in 1999 and 2004, respectively. His research interest covers machine learning, pattern recognition, data mining, and medical image analysis. Corresponding author of this paper.)