基于超图正则化的多模态信息融合算法

崔冰晶，张懿璞，王飚

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900144

基于超图正则化的多模态信息融合算法

崔冰晶　张懿璞　王　飚
长安大学电子与控制工程学院　西安 710061
(2020132049@chd.edu.cn)

摘　要　多模态数据融合方法通过学习多个数据集间的关联信息和互补信息,提高了数据分类或预测的性能。但现有的数据融合方法大都基于单独数据集自身的特征模式进行学习,不同异构数据之间的结构信息往往被忽略。因此,文中提出了一种基于超图正则化的多模态信息融合算法(sHMF),通过超图和流行正则项的方法结合表示模态内样本间的高阶关系和模态间的关系,即得到同构和异构的高阶网络。其中,采用超图稀疏表达学习超图,减少冗余边。为了验证所提算法的性能,在模拟数据和影响遗传学真实数据下进行实验,结果表明,sHMF 算法在模拟数据和真实数据上均优于多任务学习、多邻域分类等流行算法对精神分裂症的分类精度。同时,sHMF 在真实数据上得出的实验结果进一步揭示了一些与精神分裂症显著相关的生物标记物以及风险基因、甲基化因子和异常脑区之间潜在的联系。

关键词:多模态数据融合;流形正则化;功能磁共振;影像遗传学
中图法分类号　TP391

Multimodal Data Fusion Algorithm Based on Hypergraph Regularization

CUI Bingjing,ZHANG Yipu and WANG Biao
School of Electronic and Control Engineering,Chang'an University,Xi'an 710061,China

Abstract　The multi-modal data fusion improves the performance of data classification and prediction by learning the correlation information and complementary information between multiple datasets. However,existing data fusion methods are based on feature pattern learning of single dataset and ignore structural information among different heterogeneous datasets. This paper proposes a multi-modal data fusion algorithm based on hypergraph regularization(sHMF),acquiring hyper-order relationship of inter-and cross-modality by combining hypergraph and manifold regularization,i.e. homogeneous and heterogeneous high-order networks. Specifically,it firstly generates a hypergraph similarity matrix to represent the high-order relationships among subjects. In the proposed method,the sparse representation of hypergraph is used to build hypergraph for reducing redundant hype-redges. sHMF is validated on the simulated data and real imaging genetic data of schizophrenia patients. Experiment results show that our algorithm outperforms several widely used methods in the classification accuracy of simulated data and real data,and reveals some biomarkers significantly associated with schizophrenia and the potential links between risk genes,methylation factors and abnormal brain regions.

Keywords　Multi-modal data fusion,Manifold regularization,Functional magnetic resonance imaging,Imaging genetics

1　引言

随着信息获取技术的飞速发展,多模态数据融合技术受到了人们的广泛关注。这种技术利用不同模态间相互补充的信息,被广泛应用于传感器网络、视频处理、医学成像和智能系统设计等领域[1]。影响遗传学作为一类典型的多模态数据研究,其数据具有高维度、小样本的特点,还包含多样性和异构性。近年来,为了探索不同模态间的共有特征以表达其相关性,研究者们提出了一些多模态数据融合的算法。但此类方法大都基于特征学习,并没有充分利用异构数据之间的结构信息。因此,如何利用多模态数据的结构信息进行数据融合还需要进一步的研究。

最新研究表明,融合多种影像学和遗传学数据可以提高对疾病机制的理解和临床诊断水平。例如,Adeli 等提出了基于不完整纵向影像数据的多任务多元线性回归模型,用于预测多重认知评分[2]。Jie 等提出了一种基于超图的多任务特征选择方法,用于对阿尔茨海默病和轻度认知障碍分类[3]。Zhu 等通过一种多模态多任务学习模型,联合选择少量的共同特征,并利用支持向量机(SVM)对特征进行融合、分类及回归[4]。Du 等提出了一种新的多任务 SCCA (MTSCCA)

到稿日期:2022-09-15　返修日期:2022-12-09
基金项目:国家重点研发计划(2021YFB1600200,2021YFB2601300)
This work was supported by the National Key R&D Program of China(2021YFB1600200,2021YFB2601300).
通信作者:张懿璞(zyipu@chd.edu.cn)

方法，用于识别单核苷酸多态性（SNP）数据和多模态成像数量性状之间的双多变量关联[5]。Bai 等将基于图的半监督学习（GSSL）模型应用于精神疾病分类[6]。Hu 等开发了一种基于可解释深度网络的多模态融合模型，可自动诊断和对结果进行解释，用于脑影像遗传学数据研究[7]。

与传统单一观测数据相比，多模态影像遗传学数据具有高维度、小样本的特点，且每个模态都相互独立且服从于不同分布。目前的研究大多先对每个模态的数据降维，再利用相似度评分来寻找模态间的关联性，不能反映多模态数据间复杂的相互作用。为了解决上述问题，本文提出了一种基于超图正则化的多模态数据融合算法（sHMF），用于融合多种影像和遗传数据。首先引出了一种超图稀疏学习模型，生成超图相似矩阵表示样本间的高阶关系。考虑到模态间的关系，在此基础上通过流行正则化，提出了基于超图正则化的多模态数据融合算法（sHMF），既保留了模态内样本间的关系又增加了模态间的关系，得到了同构和异构的高阶网络。为验证 sHMF 算法的有效性，实验分别采用了模拟数据集和精神分裂症的真实数据。模拟实验结果表明，sHMF 具有较强的鲁棒性，能识别多模态数据的重要特征，分类精度高于其他算法。真实实验结果表明，sHMF 算法融合多模态数据比单模态数据具有更好的分类性能。同时，sHMF 在真实数据上得出的实验结果进一步揭示了一些与精神分裂症显著相关的生物标记物以及风险基因、甲基化因子和异常脑区之间潜在的联系。

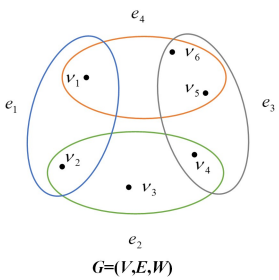
2 多模态数据融合算法

2.1 超图

普通图表示两个顶点间的关系，通过边连接两个顶点，每条边的权重表示两个顶点间的某种关系。实际中，数据间的关系比普通图表示的二阶关系更为复杂，为避免丢失普通图无法发现的有用信息，提出了超图[8]。超图可描述多个顶点间的高阶关系，通过超边连接多个顶点，其中每个超边可以连接任意数量的顶点，即超边是顶点的非空子集。

图 1 中表示超图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E},\boldsymbol{W})$ 和超图 $\boldsymbol{G}$ 的关联矩阵是 $\boldsymbol{H}$。其中超图 $\boldsymbol{G}$ 由顶点集 $\boldsymbol{V}=\{v_1,v_2,\cdots,v_6\}$、超边集 $\boldsymbol{E}=\{e_1,e_2,e_3,e_4\}$、超边权重集 $\boldsymbol{W}=[w(e_1),w(e_2),w(e_3),w(e_4)]$ 构成。关联矩阵 $\boldsymbol{H}=[H_{nk}]\in\mathbb{R}^{N\times K}$，其中 N 表示顶点个数，K 表示超边的个数，可以表示为：

$$H_{nk}=\begin{cases}1, & \text{if } v_n\in e_k\\ 0, & \text{otherwise}\end{cases} \tag{1}$$



	e_1	e_2	e_3	e_4
v_1	1	0	0	1
v_2	1	1	0	0
v_3	0	1	0	0
v_4	0	1	1	0
v_5	0	0	1	1
v_6	0	0	1	1

H

图 1 超图的表示 $\boldsymbol{G}$ 和关联矩阵 $\boldsymbol{H}$

Fig. 1 Hypergraph representation $\boldsymbol{G}$ and correlation matrix $\boldsymbol{H}$

顶点度和超边度分别定义为：

$$d(v_n)=\sum_{e_k\in E}w(e_k)H_{nk}\ \text{ for } 1\leqslant n\leqslant N \tag{2}$$

$$\phi(e_k)=\sum_{v_n\in V}H_{nk}\ \text{ for } 1\leqslant k\leqslant K \tag{3}$$

采用对角矩阵表示顶点度、超边度和超边权重矩阵：$\boldsymbol{D}_v=diag(d(v_1),d(v_2),\cdots,d(v_N))\in\mathbb{R}^{N\times N}$，$\boldsymbol{D}_e=diag(\delta(e_1),\delta(e_2),\cdots,\delta(e_K))\in\mathbb{R}^{K\times K}$，$\boldsymbol{W}=diag(w(e_1),w(e_2),\cdots,w(e_K))\in\mathbb{R}^{K\times K}$。超图中，通过计算两个顶点 i 和 j 在所有超边的权重占比来表示它们之间的关系：

$$S_{ij}=\sum_{e_k\in E}\frac{w(e_k)}{\delta(e_k)}H_{ik}H_{jk} \tag{4}$$

推导出超图 $\boldsymbol{G}$ 的相似矩阵 $\boldsymbol{S}$：

$$S=\boldsymbol{H}\boldsymbol{W}\boldsymbol{D}_e^{-1}\boldsymbol{H}^{\mathrm{T}} \tag{5}$$

通过传统图对拉普拉斯矩阵的定义，超图拉普拉斯矩阵可定义为：

$$\boldsymbol{L}=\boldsymbol{D}_v-\boldsymbol{S} \tag{6}$$

为了得到相似矩阵 $\boldsymbol{S}$，由式(5)知相似矩阵 $\boldsymbol{S}$ 与关联矩阵 $\boldsymbol{H}$ 和超边权重矩阵 $\boldsymbol{W}$ 有关。$\boldsymbol{R}=(r_1,r_2,\cdots,r_N)^{\mathrm{T}}\in\mathbb{R}^N$ 是定义在超图上的任何信号，向量 r_n 是每个顶点 v_n 的值，并对超图信号的变化进行平滑表示：

$$\begin{aligned}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{R}&=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}S_{ij}(r_i-r_j)^2\\&=\frac{1}{2}\sum_{e_k\in\boldsymbol{E}}\sum_{v_i,v_j\in\boldsymbol{V}}\frac{w(e_k)H_{ik}H_{jk}}{\delta(e_k)}(r_i-r_j)^2\end{aligned} \tag{7}$$

2.2 同构模态内关系的超图表示

假设 M 个模态的数据集中，第 m 个模态表示为 $\boldsymbol{R}_m=[r_m^1,r_m^2,\cdots,r_m^N]^{\mathrm{T}}\in\mathbb{R}^{N\times F_m}$ $(m=1,2,\cdots,M)$，其中 N 表示样本数，F_m 表示第 m 个模态样本的特征数，超图中将每个样本看成一个顶点。以下表示均为第 m 个模态的相关变量或矩阵，超图相似矩阵 $\boldsymbol{S}_m=[S_m^{ij}]\in\mathbb{R}^{N\times N}$。$\boldsymbol{T}=[t_1,t_2,\cdots,t_N]^{\mathrm{T}}$ 为样本数据的响应向量，$t_n\in\{0,1\}$ 是第 n 个样本的类别标签。

首先，通过稀疏线性回归[9]，得到第 m 个模态的关联矩阵 $\boldsymbol{H}_m$，每个样本由其他 $N-1$ 个样本线性组合表示：

$$\min_{\zeta_m^n}\frac{1}{2}\|r_m^n-\boldsymbol{P}_m^n\boldsymbol{\zeta}_m^n\|_2^2+\boldsymbol{\tau}_m\|\boldsymbol{\zeta}_m^n\|_1 \tag{8}$$

其中，$\boldsymbol{P}_m^n=(r_m^1,r_m^2,\cdots,r_m^{n-1},r_m^{n+1},\cdots,r_m^N)\in\mathbb{R}^{F_m\times(N-1)}$ 是除第 n 个样本之外其他的所有样本。$\boldsymbol{\zeta}_m^n\in\mathbb{R}^{(N-1)}$ 是系数向量，量化其他样本对第 n 个样本的影响程度。$\tau_m>0$ 是正则化参数，用于控制稀疏性。在第 m 个模态中，超边 e_n 是由第 n 个样本与系数 $\boldsymbol{\zeta}_m^n$ 是正值的 $N-1$ 种其他样本组成，得到每个样本和其他样本在同一个超边的关系。因对第 m 个模态中每个样本进行稀疏回归，故有 N 个超边且关联矩阵 $\boldsymbol{H}_m\in R^{N\times N}$ 随 τ_m 的增加或减小变得稀疏或紧密。关联矩阵 $\boldsymbol{H}\in R^{N\times NM}$ 由各个模态的关联矩阵组成，即 $\boldsymbol{H}=(\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots,\boldsymbol{H}_M)$。获得关联矩阵 $\boldsymbol{H}$ 后，通过多个超图学习获得权重向量 $\boldsymbol{w}$，原理是模态中每个特征下所有样本（即 $\boldsymbol{R}_m$ 的每一列）的观察值在超图上是平滑的，其表达式如下：

$$\begin{aligned}&\min_{\boldsymbol{w}}\frac{1}{M}\sum_{m=1}^{M}\mathrm{tr}(\boldsymbol{R}_m^{\mathrm{T}}\boldsymbol{L}_m\boldsymbol{R}_m)+\rho\|\boldsymbol{w}\|_2^2\\&\text{s.t. }\boldsymbol{w}\geqslant 0,\|\boldsymbol{w}\|_1=1\end{aligned} \tag{9}$$

其中，$\boldsymbol{w}\in\mathbb{R}^{NM}$ 是由所有 $\boldsymbol{w}_m\in\mathbb{R}^N$ $(m=1,2,\cdots,M)$ 连接组成，

即 $\boldsymbol{w}=(w_1;w_2;\cdots;w_M)$。$\boldsymbol{L}_m$ 为第 m 个模态的超图拉普拉斯矩阵。得到所有模态的关联矩阵 $\boldsymbol{H}$ 和权重向量 $\boldsymbol{w}$ 后，计算所有模态的超图相似矩阵：

$$\boldsymbol{S}=\boldsymbol{HWD}_e^{-1}\boldsymbol{H}^{\mathrm{T}}=\sum_{m=1}^{M}\boldsymbol{S}_m \tag{10}$$

其中，$\boldsymbol{W}=diag(w)$，$\boldsymbol{D}_e=diag(\boldsymbol{D}_{e1},\boldsymbol{D}_{e2},\cdots,\boldsymbol{D}_{eM})$ 且 $\boldsymbol{S}_m=\boldsymbol{H}_m\boldsymbol{W}_m(\boldsymbol{D}_{em})^{-1}(H_m)^{\mathrm{T}}$ 是第 m 个模态的超图相似矩阵。同时，为了验证超图对结构稀疏的影响，基于超图的高阶结构引入流行正则化。

2.3 流形正则项构建

为保留数据的基本结构和相似性信息，在建立单一模态内同构关系的同时建立模态间的异构关系，本文提出利用流形正则化的多模态数据融合算法，将单一模态内的样本关系和不同模态间的关系融合，正则项表达式为：

$$\boldsymbol{\pi}(\boldsymbol{\alpha},\gamma,\varphi)=\frac{1}{2}\sum_{p,q}^{M}\sum_{i,j}^{N}\theta_{p,q}S_{p,q}^{i,j}(\hat{t}_p^i-\hat{t}_q^j)^2 \tag{11}$$

正则项 $\pi(\alpha,\gamma,\varphi)$ 由两部分组成，当 $p=q$ 时，$\theta_{p,q}=\gamma$；当 $p\neq q$ 时，$\theta_{p,q}=\varphi$。第一部分如下：

$$\gamma\frac{1}{2}\sum_{p,q}^{M}\sum_{i,j}^{N}S_p^{i,j}(\hat{t}_p^i-\hat{t}_q^j)^2 \tag{12}$$

其表示单个模态内样本之间的关系。第二部分如下：

$$\varphi\frac{1}{2}\sum_{p,q}^{M}\sum_{i,j}^{N}S_{p,q}^{i,j}(\hat{t}_p^i-\hat{t}_q^j)^2 \tag{13}$$

其表示不同模态间样本的相似关系。其中 γ 和 φ 是两个部分的参数，$\hat{t}_p^i$ 是第 p 个模态中第 i 个样本的标签，$S_p^{i,j}$ 表示同一模态 p 下第 i 个样本和第 j 个样本的相似度，$S_{p,q}^{i,j}$ 表示第 p 个模态中第 i 个样本和第 q 个模态中第 j 个样本的相似度。根据文献[10]，$S_{p,q}^{i,j}$ 可以通过估计第 i 个样本与第 p 个模态中的样本，以及第 j 个样本与第 q 个模态中样本的关系来计算：

$$S_{p,q}^{i,j}=\sum_{n=1}^{N}S_p^{in}S_q^{nj} \tag{14}$$

其矩阵形式为 $\boldsymbol{S}=\boldsymbol{S}_p\boldsymbol{S}_q$，其中 $\boldsymbol{S}_p$ 和 $\boldsymbol{S}_q$ 是第 p 和 q 个模态的相似矩阵。将上述两部分相结合，得到分块相似矩阵 $\widetilde{\boldsymbol{S}}\in\mathbb{R}^{NM\times NM}$：

$$\widetilde{\boldsymbol{S}}=\begin{bmatrix}\boldsymbol{S}_{1,1}&\cdots&\boldsymbol{S}_{1,M}\\ \vdots&\ddots&\vdots\\ \boldsymbol{S}_{M,1}&\cdots&\boldsymbol{S}_{M,M}\end{bmatrix} \tag{15}$$

然后，计算拉普拉斯矩阵 $\widetilde{\boldsymbol{L}}=\widetilde{\boldsymbol{D}}-\widetilde{\boldsymbol{S}}$。因为 $\widetilde{\boldsymbol{S}}$ 是对称阵，所以 $\widetilde{\boldsymbol{D}}^{ii}=\sum\widetilde{\boldsymbol{S}}^{ij}$，$1\leqslant i\leqslant NM$。结合正则化参数 γ 和 φ，矩阵的分块形式 $\widetilde{\boldsymbol{L}}$ 为：

$$\widetilde{\boldsymbol{L}}=\begin{bmatrix}\gamma\boldsymbol{L}_{1,1}&\cdots&\varphi\boldsymbol{L}_{1,M}\\ \vdots&\ddots&\vdots\\ \varphi\boldsymbol{L}_{M,1}&\cdots&\gamma\boldsymbol{L}_{M,M}\end{bmatrix} \tag{16}$$

通过式(16)可以等价地表示为：

$$\pi(\boldsymbol{\alpha},\gamma,\varphi)=\begin{bmatrix}\boldsymbol{R}_1\boldsymbol{\alpha}_1\\ \boldsymbol{R}_2\boldsymbol{\alpha}_2\\ \cdots\\ \boldsymbol{R}_M\boldsymbol{\alpha}_M\end{bmatrix}^{\mathrm{T}}\widetilde{\boldsymbol{L}}\begin{bmatrix}\boldsymbol{R}_1\boldsymbol{\alpha}_1\\ \boldsymbol{R}_2\boldsymbol{\alpha}_2\\ \cdots\\ \boldsymbol{R}_M\boldsymbol{\alpha}_M\end{bmatrix} \tag{17}$$

本文的 sHMF 模型为：

$$\Gamma=\min_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_m}\frac{1}{2}\sum_{m=1}^{M}[\|\boldsymbol{T}-\boldsymbol{R}_m\boldsymbol{\alpha}_m\|_2^2+\lambda_m\|\boldsymbol{\alpha}_m\|_1]+\pi(\alpha,\gamma,\varphi) \tag{18}$$

其中，γ 和 φ 表示流形正则项的参数，用于调整模态内和模态间的样本相似性。图 2 为 sHMF 算法的框架图。

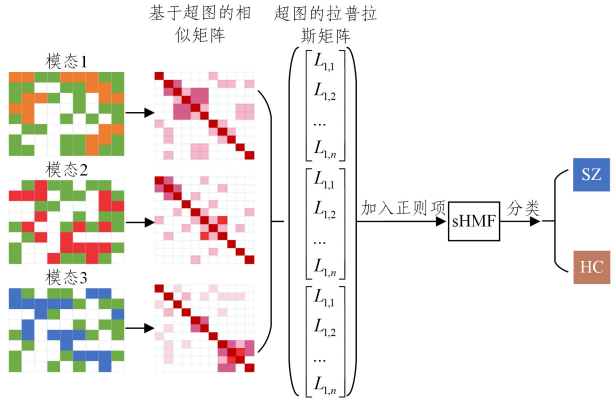


图 2 sHMF 算法的分析框架

Fig. 2 Analytical framework of sHMF algorithm

式(18)中的目标函数是凸函数，为了优化 $\boldsymbol{\alpha}$，对 $\boldsymbol{\alpha}_m$ 求导：

$$\nabla\Gamma(\boldsymbol{\alpha}_m)=(\boldsymbol{R}_m)^{\mathrm{T}}(\boldsymbol{R}_m\boldsymbol{\alpha}_m-\boldsymbol{T})+\gamma\sum_{m=1}^{M}(\boldsymbol{R}_m)^{\mathrm{T}}\boldsymbol{L}_{m,m}\boldsymbol{\alpha}_m+\varphi\sum_{m=1}^{M}\sum_{p=1,(p\neq m)}((\boldsymbol{R}_m)^{\mathrm{T}}\boldsymbol{L}_{p,m}\boldsymbol{R}_p\boldsymbol{\alpha}_p)+\lambda_m\mathrm{sgn}(\boldsymbol{\alpha}_p) \tag{19}$$

其中，$\mathrm{sgn}(\cdot)$ 为符号函数。$\boldsymbol{\alpha}_m$ 的迭代更新公式如下：

$$\boldsymbol{\alpha}_m^{k+1}=((\boldsymbol{R}_m)^{\mathrm{T}}\boldsymbol{R}_m+\gamma(\boldsymbol{R}_m)^{\mathrm{T}}\boldsymbol{L}_{m,m}\boldsymbol{R}_m)\times((\boldsymbol{R}_m)^{\mathrm{T}}-\varphi\sum_{p=1}^{M}\sum_{q=1}^{M}((\boldsymbol{R}_m)^{\mathrm{T}}\boldsymbol{L}_{p,q}\boldsymbol{\alpha}_p^k))+\lambda_m\mathrm{sgn}(\boldsymbol{\alpha}_m^k)(p\neq q,p\parallel q=m) \tag{20}$$

其中，k 为第 k 次迭代。当目标函数的相对误差满足 $|\Gamma(k+1)-\Gamma(k)|/|\Gamma(k+1)|\leqslant10^{-6}$ 时，迭代结束。算法 1 为 sHMF 算法的伪代码。

算法 1 基于超图正则化的多模态信息融合算法

输入：M 个模态数据 $\mathbf{R}_m$，响应向量 $\mathbf{T}$，超图参数 τ_m，ρ，γ，φ 和 λ_m（m=1,2,…,M）

输出：权重向量 $\boldsymbol{\alpha}_m$

1. 初始化：$\boldsymbol{\alpha}_m$
2. for m=1 to M do
3. 获取每个数据集的 $\mathbf{H}_m$ 和 $\boldsymbol{\alpha}_m$
4. 通过式(10)计算 $\mathbf{S}_m$
5. end for
6. 通过 $\mathbf{S}_m$ 得到 $\widetilde{\mathrm{S}}$，计算 $\widetilde{\mathbf{L}}$
7. for k=1 to Max-Iteration do
8. 通过式(20)计算 $\boldsymbol{\alpha}$
9. until converge
10. end for
11. Return $\boldsymbol{\alpha}_m$(m=1,2,…,M)

3 实验结果与分析

3.1 模拟数据实验

3.1.1 模拟数据生成

为了评估 sHMF 算法的性能，本文首先模拟 3 种不同属性的数据集。为模拟高维度、小样本数据特点，样本的数量远小于特征的数量。$\boldsymbol{B}_1\in\mathbb{R}^{f\times n_1}$，$\boldsymbol{B}_2\in\mathbb{R}^{f\times n_2}$ 和 $\boldsymbol{B}_3\in\mathbb{R}^{f\times n_3}$，其中 $f=200$ 是样本个数，$n_1=3\,000$，$n_2=4\,000$ 和 $n_3=5\,000$ 分别是

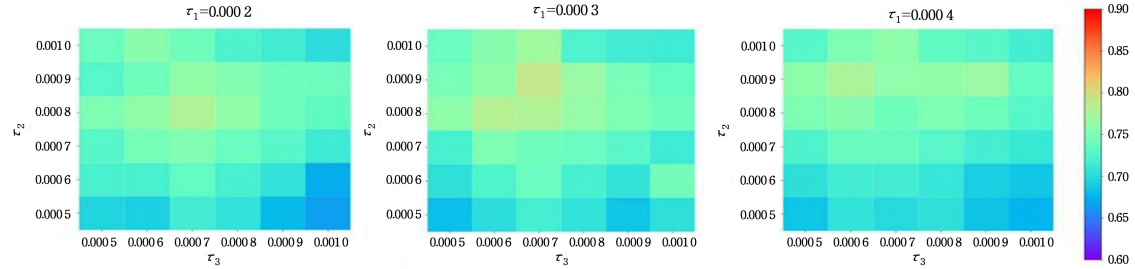
3 个数据集的特征维度。模拟过程与文献[11]相似，具体步骤如下：首先假设 4 个独立的潜在变量 $c_1, c_2, c_3, c_4 \in \mathbb{R}^{200\times 1}$ 服从标准正态分布，其次创造 3 个稀疏变量 $\boldsymbol{\mu} \in \mathbb{R}^{1\times n_i}$，每个数据集可以形成如下形式：

$$\begin{aligned}\boldsymbol{B}_1 &= c_1\ \boldsymbol{\mu}_1 + c_2\ \boldsymbol{\mu}_2 + c_4\ \boldsymbol{\mu}_3 + \boldsymbol{Z}_1\\ \boldsymbol{B}_2 &= c_1\ \boldsymbol{\mu}_4 + c_3\ \boldsymbol{\mu}_5 + c_4\ \boldsymbol{\mu}_6 + \boldsymbol{Z}_2\\ \boldsymbol{B}_3 &= c_2\ \boldsymbol{\mu}_7 + c_3\ \boldsymbol{\mu}_8 + c_4\ \boldsymbol{\mu}_9 + \boldsymbol{Z}_3\end{aligned} \tag{21}$$

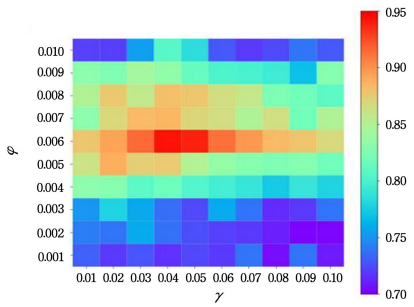
其中，$\boldsymbol{Z}_1 \in \mathbb{R}^{120\times 3000}$，$\boldsymbol{Z}_2 \in \mathbb{R}^{120\times 4000}$ 和 $\boldsymbol{Z}_3 \in \mathbb{R}^{120\times 5000}$ 服从标准正态分布，每个向量 $\boldsymbol{\mu}$ 包含 100 个非零元素，每一个都服从均匀分布 $\boldsymbol{\mu}_{\text{non-zeros}} \sim U(0.4, 0.6)$。$\boldsymbol{\mu}_3$，$\boldsymbol{\mu}_6$ 和 $\boldsymbol{\mu}_9$ 是前 100 个特征，$\boldsymbol{\mu}_1$ 和 $\boldsymbol{\mu}_4$ 是第 101－200 个特征，$\boldsymbol{\mu}_2$ 和 $\boldsymbol{\mu}_7$ 是第 201－300 个特征，$\boldsymbol{\mu}_5$ 和 $\boldsymbol{\mu}_8$ 是第 301－400 个特征。因此，$\boldsymbol{B}_1$ 和 $\boldsymbol{B}_2$ 具有潜在变量 c_1 产生的 100 个相关特征，$\boldsymbol{B}_1$ 和 $\boldsymbol{B}_3$ 具有潜在变量 c_2 产生的 100 个相关特征，$\boldsymbol{B}_2$ 和 $\boldsymbol{B}_3$ 具有由潜在变量 c_3 产生的 100 个相关特征，$\boldsymbol{B}_1$，$\boldsymbol{B}_2$ 和 $\boldsymbol{B}_3$ 具有由潜在变量 c_4 产生的 100 个相关特征。这些解释变量生成连续标签，将其转换为 Z 分数。从每个数据集两端分别选取得分在前 30%和后 30%的样本，保留 $f=200$ 个样本，并根据分数分配二进制标签。最后，在 3 个数据集中加入噪声信号，验证本文算法的鲁棒性。

3.1.2 参数选择

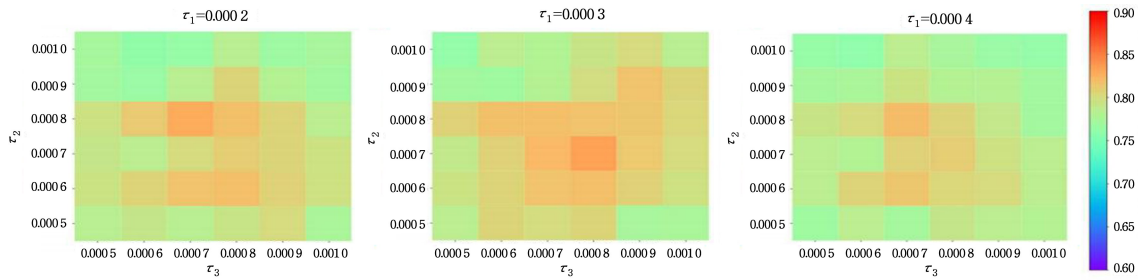
在模拟数据实验中，本文使用 10 折交叉验证来评估预测模型的分类精度。先将数据集随机划分为 10 个不相交的子集，其中一个子集作为测试集，其余 9 个子集用于训练。此过程重复 10 次，可减少抽样偏差的影响。通过网格搜索对模型正则化参数进行调整，其中包括流行正则化参数 γ 和 φ，模型稀疏参数 $\lambda_1, \lambda_2, \lambda_3$ 和超图稀疏参数 τ_1, τ_2, τ_3 以及超图正则化参数 ρ。为了减少实验次数，本文将参数调优过程分为 4 部分，如图 3 所示。



(a)sHMF 模型超图系数参数的选择



(b)sHMF 模型正则化参数的选择



(c)sHMF 模型稀疏参数的选择

图 3 本文算法在不同参数设置下的分类性能

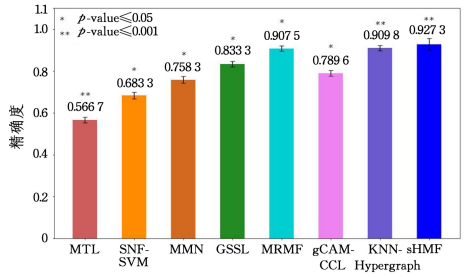
Fig. 3 Classification performance of the proposed algorithm with respect to different parameters' settings

图 3(a)为没有稀疏约束的参数调优，即当 $\gamma, \varphi, \rho=0$ 以及 $\lambda_1, \lambda_2, \lambda_3=0$ 时，超图稀疏参数 $\tau_1, \tau_2, \tau_3 \in \{1\times 10^{-2}, 2\times 10^{-2}, \cdots, 1\times 10^{-1}\}$ 在各自范围内选取的部分热力图。经分析得 $\tau_1=0.03, \tau_2=0.05, \tau_3=0.05$ 时模型的精度最高。由图 3(b)可知，得到超图稀疏参数后，超图正则化参数 $\rho \in \{1\times 10^{-3}, 2\times 10^{-3}, \cdots, 1\times 10^{-2}\}$ 在各自范围内选取，经分析确定 $\rho=0.004$ 时模型的精度最高。由图 3(c)可知，当 $\tau_1=0.03$，$\tau_2=0.05, \tau_3=0.05, \rho=0.004$ 时，流行正则化参数 $\gamma \in \{1\times 10^{-2}, 2\times 10^{-2}, 3\times 10^{-2}, \cdots, 1\times 10^{-1}\}$ 和 $\varphi \in \{1\times 10^{-3}, 2\times 10^{-3}, 3\times 10^{-3}, \cdots, 1\times 10^{-2}\}$ 在各自范围内选取参数的部分热力图。最终确定 $\tau_1=0.03, \tau_2=0.05, \tau_3=0.05, \rho=0.004, \gamma=0.02, \varphi=0.006$ 时模型的精度最高。图 3(c)为参数取图 3(a)中的参数时，稀疏参数 $\lambda_1, \lambda_2 \in \{1\times 10^{-4}, 2\times 10^{-4}, \cdots, 6\times 10^{-4}\}$，$\lambda_3 \in \{3\times 10^{-3}, 4\times 10^{-3}, \cdots, 8\times 10^{-3}\}$ 在各自范围内选取的部分热力图。模型精度因参数改变会发生明显变化，不断调整参数大小，最终将参数设定为：$\tau_1=0.03, \tau_2=0.05,$

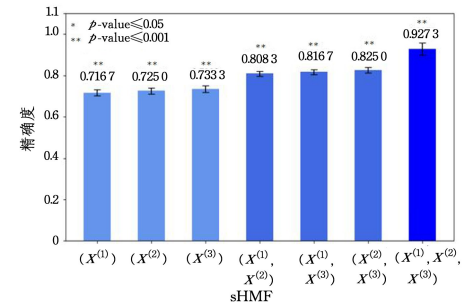
$\tau_3=0.05$，$\rho=0.004$，$\gamma=0.02$，$\varphi=0.006$，$\lambda_1=0.002$，$\lambda_2=0.0004$，$\lambda_3=0.0005$ 时达到了最高的分类准确率，即 0.9273。

3.1.3 模拟数据实验结果

在模拟数据实验中，比较 sHMF 模型与其他 7 个模型的分类性能，例如基于 KNN 构建超图的多模态融合[11]（KNN-Hypergraph）、多任务学习（MTL）[12]、基于相似度网络融合的 SVM（SNF-SVM）[13]、基于均值融合的多邻域分类（MMN）[14]、基于图的半监督学习（GSSL）[15]、梯度 CAM 引导的卷积协同学习（gCAM-CCL）[7]、基于图网络的流行正则化多模态数据融合（MRMF）。同时，通过组合不同数据集测试 sHMF 算法，结果如图 4 所示。



（a）sHMF 在不同的数据集中的分类精度



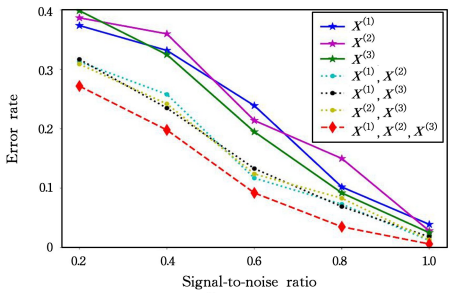
（b）sHMF 与其他 6 种算法的分类精度比较

图 4 模拟数据集中不同数据组合和不同算法的分类性能

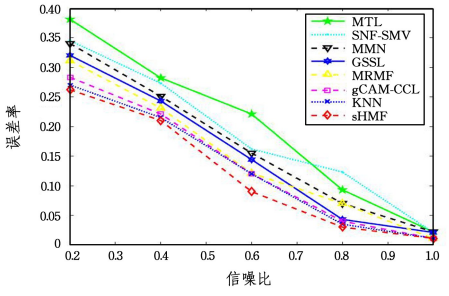
Fig. 4 Classification performance of different combination of data and different algorithms on simulated data sets

图 4(a)中，sHMF 模型通过对融合的 3 组数据集分类，获得最佳分类精度 0.9273±0.0293。图 4(b)中，sHMF 在单个数据集 B_1，B_2，B_3 上的分类精度分别为 0.7167±0.0137，0.7250±0.0154，0.7333±0.0163；在两个数据集上的分类精度高于单个数据集的分类精度，两个数据集 B_1 和 B_2，B_1 和 B_3，B_2 和 B_3 的分类精度为 0.8083±0.0133，0.8167±0.0129，0.8250±0.0128；在 3 个数据集 B_1，B_2 和 B_3 上的分类精度为 0.9273±0.0293。这说明多模态数据融合相比单一数据更具优势且来源广泛，更能提取有用信息。

为了进一步验证算法，在不同噪声水平（即信噪比，SNR）的模型数据中比较 sHMF 在不同数据集上的效果，信噪比∈[0.2，0.4，0.6，0.8，1.0]，如图 5 所示（为了图 5 的简洁性，KNN-Hypergraph 用 KNN 简写表示）。图 5(a)给出了在 5 种不同信噪比下在不同数据集上 sHMF 测试的错误率。经分析得，多个数据相比单一数据的稳定性更好，精度更高。同时，在不同噪声水平的模拟数据上比较 sHMF 和其他算法，图 5(b)说明 sHMF 算法在不同噪声水平下比其他算法的误差率低，适用于不同场景，在不同的复杂情况下保持了较好的模型性能。



（a）sHMF 在不同数据集组合和不同噪声水平中的误差率



（b）sHMF 与其他算法在不同噪声水平下的误差率

图 5 不同噪声水平对数据组合和算法的影响

Fig. 5 Influence of different SNRs on datasets combination and algorithms

实验结果表明，利用多模态数据确实可以提高分类精度，并且比使用单一数据集或合并两个数据集具有更好的抗噪性。与其他算法相比，本文算法中基于超图的流形正则项能够有效地利用多模态数据的同质和异质结构，并为解决高维度、小样本数据的过拟合问题提供了一种有效途径。

3.2 真实数据实验

3.2.1 真实数据的准备和预处理

在真实数据实验中，本文利用精神分裂症患者和健康对照组的真实数据对 sHMF 算法进行测试。这些真实样本包含精神临床影像联盟（MCIC）收集的 SNP、DNA 甲基化和 fMRI 数据集[15]。MCIC 数据集有 183 例受试者，其中包含 79 例精神分裂症患者（年龄为 34±11，女性 20 例）和 104 例健康对照组（年龄为 32±11，女性 38 例）。

fMRI 数据是在受试者执行感觉运动任务时收集的，这是一种对听觉刺激运动反应的分组实验。fMRI 数据使用 SPM 软件进行预处理，其过程包括重新对齐、空间归一化和平滑在内的连续处理。因音频刺激，本文通过多元回归分析数据[16]。从每个受试者中提取一个 53×63×46 的刺激对比图像，排除缺少测量值的体素后，每个图像共包含 41 236 个体素，根据自动解剖标记 AAL（Anatomical Automatic Labeling）模板[17]将其划分为 116 个大脑感兴趣区域（ROI）。

SNP 数据是从每个受试者的血液样本中获取的，每个受试者均采集血样并提取 DNA。在 Mind 研究网络中使用 Illumina Infinium HumanOmni1-Quad 软件对所有受试者进行基因分型[18]，覆盖 1 140 419 个 SNP 位。用 PLINK 软件来执行一系列标准的质量控制程序，最终得到涵盖 722 177 个基因座的数据集。

DNA 甲基化数据也从血液样本中提取，并通过 Illumina Infinium methylation 27 k Assay 实验进行评估。用 R pa-

ckage watermelon 归一化后，数据包含 27 508 CpG 位点信息。每个量介于 0 和 1 之间，代表每个 CpG 位点的甲基化水平[19-20]，在去除方差小于 1×10^{-4} 的位点后进一步选择 9 273 个甲基化位点。

3.2.2 真实数据的实验结果

为了进一步验证模型性能，本文通过 MCIC 数据集测试 sHMF 算法，并与 MTL，SNG-SVM，MMN，gCAM-CCL，MRMF 和 GSSL 算法进行比较。采用 10 折交叉验证评估这些方法的分类性能，在 183 名受试者中独立运行 10 次。用测试集中预测的受试者标签和实际的受试者标签之间的均方根误差和分类精度量化模型性能。sHMF 在真实数据实验中的最终参数设定为：$\tau_1=0.002$，$\tau_2=0.004$，$\tau_3=0.005$，$\rho=0.00006$，$\gamma=0.03$，$\varphi=0.006$，$\lambda_1=0.000005$，$\lambda_2=0.000003$，$\lambda_3=0.00003$。

sHMF 在提高分类准确性的同时通过整合多模态的高阶信息降低了均方根误差。其中，分类性能较好的算法如下：sHMF 的分类准确度为 0.874 9±0.016 4，均方根误差为 0.690 2±0.105 3；将 MRMF 看作 sHMF 的普通图，其准确度和均方根误差为 0.847 9±0.016 2，0.734 2±0.012 6；GSSL 的准确度和均方根误差为 0.820 3±0.021 7，0.788 4±0.157 3。

为测试模型性能，在 MCIC 数据集上绘制所有比较算法的 ROC 曲线，如图 6 所示（为了图 6 的简洁性，KNN-Hypergraph 用 KNN 简写表示）。图 6 中 sHMF 的 AUC 值高于其他算法，这说明所提出的超图正则化的多模态数据融合算法在 MCIC 数据中体现出了更好的性能。实验结果表明，sHMF 模型在模拟数据和 MCIC 数据集中都达到了较好的分类性能，准确度高于其他模型，且利用多数据集提供的互补信息有效提高了预测精度，减小了误差。sHMF 的 *AUC*（0.91）比基于 KNN 超图的多模态融合算法（*AUC*=0.89）、图网络模型 MRMF（*AUC*=0.87）和图学习模型 GSSL（*AUC*=0.84）更高，且高于其他模型的 *AUC*。

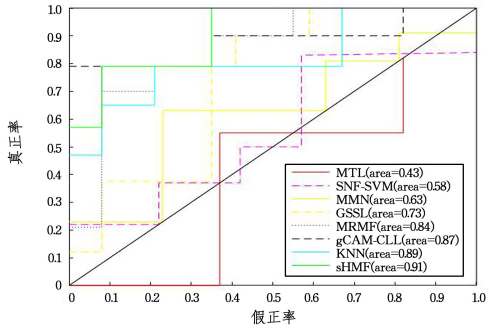


图 6 不同模型的 ROC 曲线

Fig. 6 ROC curves of 8 comparison algorithms on MCIC dataset

3.2.3 生物标记物的提取

sHMF 算法不仅在精神分裂症分类上取得了较好的性能，还获得了一些与精神分裂症显著相关的生物标记物。基于 sHMF 模型估计的权向量 $\boldsymbol{\alpha}$，研究 SNP 位点、DNA 甲基化位点和脑 ROI 分别对应的精神分裂症相关的潜在生物标志物[21-22]。

在 SNP 数据集上，sHMF 识别精神分裂症相关的前 20 个风险基因。其中，根据已有研究，CNTN2[15]，B3GATI[18]，INSIG2[23] 和 ACTG2[24] 等基因已被证实与大脑发育和神经信号有关；GWAS[25]，CSMD1 和 C20orf39 与精神分裂症具有全基因组显著相关性；增加的 mRNA 水平在基因 MAN1A1[26] 中识别，在 n-聚糖成熟反映精神分裂症中参与未成熟 n-聚糖加工多个阶段的酶基因表达失调；炎症相关基因 IL1B 在精神分裂症异常的精神运动行为特征中起着重要作用[27]，并通过 sHMF 算法发现了其他风险基因 BCAR3，AREG，RND3，LOC100131409，NCRNA00092，OLFM3，DDX26B，LOC100288860，NMNAT3，RAD23B，DMRT2，SFI1。

在 DNA 数据中，sHMF 检测到与精神分裂症相关的风险基因排名前 15 位的 DNA 甲基化表观遗传因子。其中，15q11-q13 染色体上的 TNRC4 已被证实是印痕神经发育障碍的病因，印痕神经发育障碍可能导致情感性障碍和精神病[28]；MTERF 控制线粒体转录终止功能，导致了一些精神疾病症状的临床表现[29]；DCC 可能是精神分裂症易感性个体差异的遗传基础[29]；编码 Cadherin-13（CDH13′）的基因是一种细胞粘附分子，与神经精神疾病有关[30]；PFTK1，TMEM14B，FLJ40629 与文献[31]中的精神分裂症相关，通过 sHMF 算法发现的其他风险基因为 MRPS21，TP73，IRF6′，PTAFR，IIP45，LAMB3，PLEKHA6，SFPQ。

通过 fMRI 数据分析，sHMF 算法发现了与精神分裂症相关的 ROIs。精神分裂症患者与健康对照组的差异主要集中在海马区、楔前叶和小脑[31]，精神分裂症患者的 ROIs（如 Temporal Sup_R，Insula_L 和 Lingual_L）也存在异常[15,25]。精神分裂症组左中央后回灰质（Postcentral_L）明显低于健康组[32]。文献[33]也提到了精神分裂症患者的损伤脑区比健康人大，尤其是小脑、额叶皮质、丘脑和颞叶皮质，并且额上回与 Calcarine 皮质之间的连接在显性状态下发生了特定的变化。

图 7 给出了通过 BrainNet 查看器[34] 检测到的与精神分裂症相关的 ROIs 位于 6 个脑功能区域，包括感觉-躯体运动（SM）、听觉（AUD）、默认模式（DMN）、视觉（VIS）、显著性（SN）和小脑（CER）。

综上，本文的主要贡献在于以下几个方面：1）超图不仅能表达受试者间的相互关系，而且能有效学习受试者间的高阶关系；2）考虑到模态内和模态间的相似关系，流形正则化有效表达了多模态数据的同质结构和异质结构，并且通过个体在不同模态中相互补充信息和同一模态下的相互关系，提高了模型分类和预测的精度；3）该算法结合了稀疏性和流形正则化，不仅解决了高维度、小样本数据的过拟合问题，而且增强了模型的鲁棒性。

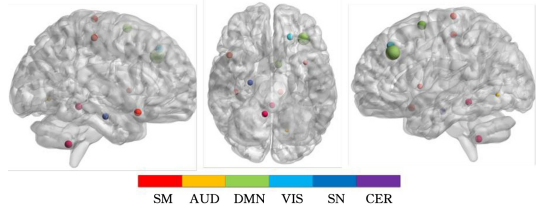


图 7 6 个异常脑域

Fig. 7 Visualization of 6 abnormal brain domains

结束语 本文基于多任务学习模型构建超图,并设计了超图正则项来实现多模态信息融合,解决了单一模态数据和两者间结构关系的问题。通过构建超图获得模态内样本间的高阶关系,将流行正则化和稀疏化相结合,获得了同构和异构高阶网络并兼顾结构信息和主体间复杂的交互作用,避免了高维低样本数据下的过拟合问题。本文提出的 sHMF 算法被应用于模拟数据集和 MCIC 数据集的精神分裂症分类研究中。实验结果表明,sHMF 算法对精神分裂症的分类性能优于其他有竞争力的算法。此外,本文在多模式数据中发现了重要的生物标志物,为探索精神分裂症的多因素发病机制提供了更多的信息。

参 考 文 献

[1] REN Z Y,WANG Z C,KE Z W,et al. Overview of multimodal data fusion [J]. Computer Engineering and Application,2021,57(18):49-64.

[2] ADELI E,MENG Y,LI G,et al. Multi-task prediction of infant cognitive scores from longitudinal incomplete neuroimaging data [J]. Neuroimage,2019,185:783-792.

[3] JIE B,ZHANG D,CHENG B,et al. Manifold regularized multi-task feature learning for multimodality disease classification[J]. Human brain mapping,2015,36(2):489-507.

[4] ZHU X,SUK H I,LEE S W,et al. Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification[J]. IEEE Transactions on Biomedical Engineering,2015,63(3):607-618.

[5] DU L,LIU K,YAO X,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.

[6] BAI Y,PASCAL Z,CALHOUN V,et al. Optimized Combination of Multiple Graphs with Application to the Integration of Brain Imaging and Genomics Data[J]. IEEE Transactions on medical imaging,2019,39(6):1801-1811.

[7] HU W,MENG X,BAI Y,et al. Interpretable multimodal fusion networks reveal mechanisms of brain cognition[J]. IEEE Transactions on medical imaging,2021,40(5):1474-1483.

[8] JI R,GAO Y,HONG R,et al. Spectral-Spatial Constraint Hyperspectral Image Classification[J]. IEEE Transactions on Geoscience and Remote Sensing,2014,52(3):1811-1824.

[9] XIAO L,WANG J,KASSANI P H,et al. Multi-Hypergraph Learning-Based Brain Functional Connectivity Analysis in fMRI Data[J]. IEEE Transactions on Medical Imaging,2020,39(5):1746-1758.

[10] XIAO L,STEPHEN J M,WILSON T W,et al. A manifold regularized multi-task learning model for IQ prediction from two fMRI paradigms[J]. IEEE Transactions on Biomedical Engineering,2019,67(3):796-806.

[11] PENG Y. Research on Multimodal Feature Selection and Classification Method Based on Hypergraph [D]. Nanjing: Nanjing University of Aeronautics and Astronautics,2019.

[12] ARGYRIOU A,EVGENIOU T,PONTIL M. Multi-Task Feature Learning[C]//Conference on Advances in Neural Information Processing Systems. Cambridge:MIT Press,2007:19-41.

[13] WANG B,MEZLINI A M,DEMIR F,et al. Similarity network fusion for aggregating data types on a genomic scale[J]. Nat Methods,2014,11(3):333-337.

[14] DENG S P,HU W,CALHOUN V D,et al. Schizophrenia prediction using integrated imaging genomic networks[J]. Advances in Science,Technology and Engineering Systems,2017,2(3):702-710.

[15] HU W,LIN D,CAO S,et al. Adaptive Sparse Multiple Canonical Correlation Analysis with Application to Imaging (Epi)Genomics Study of Schizophrenia[J]. IEEE Transactions on Biomedical Engineering,2018,65(2):390-399.

[16] TZOURIO-MAZOYER N,LANDEAU B,PAPATHANASSIOU D,et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain[J]. Neuroimage,2002,15(1):273-289.

[17] SUN L,PATEL R,LIU J,et al. Mining brain region connectivity for Alzheimer's disease study via sparse inverse covariance estimation[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris France:KDD09,2009:1335-1344.

[18] LIN 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.

[19] PIDSLEY R,Y WONG C C,VOLTA M,et al. A data-driven approach to preprocessing Illumina 450K methylation array data [J]. BMC genomics,2013,14(1):1-10.

[20] LIU J,CHEN J,EHRLICH S,et al. Methylation patterns in whole blood correlate with symptoms in schizophrenia patients [J]. Schizophrenia bulletin,2014,40(4):769-776.

[21] CHENG Y J,SUN Q Q,ZHAO M. Epigenetic research progress of methamphetamine use and addiction[J]. Shanghai:Journal of Shanghai JiaoTong University (Medical Edition),2021,41(8):1094-1098.

[22] LI M Y C,ZHANG W N,XIONG X,et al. Research progress on the balance of central excitation inhibition in an autism model with SHANK3 gene mutation [J]. Life Sciences,2021,33(8):962-970.

[23] FANG J,LIN D,SCHULZ C,et al. Joint sparse canonical correlation analysis for detecting differential imaging genetics modules[J]. Bioinformatics,2016,32(22):3480-3488.

[24] PENG P,ZHANG Y,JU Y,et al. Group Sparse Joint Non-negative Matrix Factorization on Orthogonal Subspace for Multi-modal Imaging Genetics Data Analysis[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022,19(1):479-490.

[25] WANG M,HUANG T Z,FANG J,et al. Integration of Imaging (Epi) Genomics Data for the Study of Schizophrenia Using Group Sparse Joint Nonnegative Matrix Factorization[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics,2020,17(5):1671-1681.

[26] KWON E,WANG W,TSAI L H. Validation of schizophrenia-associated genes CSMD1, C10orf26, CACNA1C and TCF4 as miR-137 targets[J]. Mol Psychiatry,2013,18(1):11-12.

[27] LEE M R, SHESKIER M B, FAROKHNIA M, et al. Oxytocin receptor mRNA expression in dorsolateral prefrontal cortex in major psychiatric disorders: A human post-mortem study[J]. Psychoneuroendocrinology, 2018, 96(1): 143-147.

[28] ZHANG Y, YOU X, LI S, et al. Peripheral Blood Leukocyte RNA-Seq Identifies a Set of Genes Related to Abnormal Psychomotor Behavior Characteristics in Patients with Schizophrenia[J]. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research, 2020, 26(1): 1-31.

[29] SAMAAN M C. Prader-Willi Syndrome: Genetics, Phenotype, and Management[J]. Current Psychiatry Reviews, 2014, 10(2): 168-181.

[30] MARAZZITI D, BARONI S, PICCHETTI M, et al. Psychiatric disorders and mitochondrial dysfunctions[J]. European Review Medical and Pharmacological Sciences, 2012, 16(2): 270-275.

[31] GRANT A, FATHALLI F, ROULEAU G, et al. Association between schizophrenia and genetic variation in DCC: A case-control study[J]. Schizophrenia Research, 2012, 137(1): 26-31.

[32] JOB D E, WHALLEY H C, MCCONNELL S, et al. Structural gray matter differences between first-episode schizophrenics and normal controls using voxel-based morphometry[J]. Neuroimage, 2002, 17(2): 880-889.

[33] DUAN H F, GAN J L, YANG J M, et al. A longitudinal study on intrinsic connectivity of hippocampus associated with positive symptom in first-episode schizophrenia[J]. Behavioral brain research, 2015, 283(1): 78-86.

[34] ZHENG J J. Research on multimodal brain network of schizophrenia based on magnetic resonance imaging[D]. Chengdu: University of Electronic Science and Technology, 2017.

CUI Bingjing, born in 1996, master candidate. Her main research interests include multi-modal data fusion and machine learning.

WANG Biao, born in 1969, Ph. D, professor. His main research interests include analysis and optimization of complex networks and systems, multi-agent control.

(责任编辑:喻藜)