计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190300073

基于深度主成分相关自编码器的多模态影像遗传数据研究

李 刚 王 超 韩德鹏 刘强伟 李 莹
长安大学电子与控制工程学院 西安 710064

摘 要 脑成像表型和基因变异已成为影响精神分裂症等复杂疾病的重要因素。研究人员根据以往在致病机理方面的深入研究,已经提出了很多基于深度神经网络或正则化的模型,这些模型通常包含某种形式的惩罚项或具有重建目标的自编码器结构,但其所使用的多模态数据的特征维数往往大于样本个数。为了应对高维数据分析的困难并突破深度典型关联分析的局限性,文中提出了一种由多模态线性特征学习的主成分分析和基于限制玻尔兹曼机的多模态非线性特征学习的多层信念网络组成的有效模型。该模型和先前的先进模型一起被应用在实际的多模态数据集上进行测试和分析。实验发现,与已有模型相比,深度主成分相关自编码器模型学习的特征具有更高的分类性能和更强的关联性。在分类精度方面,两类模态数据的分类精度均超过了90%,相比平均精度在65%左右的基于CCA的模型和平均精度在80%左右的基于DNN的模型,该模型的分类效果有了显著提高。在聚类性能评估的实验中,该模型以93.75%的平均归一化互信息指标和3.8%的平均分类错误率指标进一步验证了其优越的分类性能。在最大关联性分析方面,当顶层节点输出维度一致时,该模型以0.926的最大关联性胜于其他先进模型,在高维数据分析方面表现出了优异的性能。

关键词:影像基因组学;深度主成分相关自编码器;信念网络;优化算法;关联分析
中图法分类号 TP391

Study on Multimodal Image Genetic Data Based on Deep Principal Correlated Auto-encoders

LI Gang, WANG Chao, HAN De-peng, LIU Qiang-wei and LI Ying
School of Electronic and Control Engineering, Chang'an University, Xi'an 710064, China

Abstract Brain imaging phenotype and genetic mutation has become the important factors that affect complex diseases such as schizophrenia, researchers based on previous work in the pathogenesis of in-depth research have proposed many models based on deep neural network or regularization, typically involving either some form of norm or auto-encoders with a reconstruction objective, but the multi-modal data of those models tend to have the number of feature dimensions which more than that of samples. In order to solve the difficulties of high-dimensional data analysis and overcome the limitations of deep canonical correlation analysis, a competent optimization algorithm is exploited to solve deep canonical correlation analysis (DCCA) with principal component analysis (PCA) on the multi-modal linear features learning and multi-layer belief network based on restricted Boltzmann machine (RBM) on multi-modal nonlinear features learning. The model, together with previous advanced model, has been applied to test and analyze the actual multi-modal data. Experiments show that the deep principal component correlation auto-encoders model has higher correlation and better classification performance than those previous model. In terms of classification accuracy, the classification accuracy of the two types of modal data is more than 90%. Compared with the CCA-based model with an average accuracy of about 65% and the DNN-based model with an average accuracy of about 80%, the classification effect of this model is significantly improved. In the experiment of clustering performance evaluation, the model further verified the significant classification effect of the model with average normalized mutual information of 93.75% and average classification error rate of 3.8%. In terms of maximum correlation analysis, on the premise that the output dimensions of top-level nodes are consistent, this model outperforms other advanced models with the maximum correlation of 0.926, showing excellent performance in high-dimensional data analysis.

Keywords Image genomics, Deep principal correlated auto-encoders, Belief networks, Optimization algorithms, Correlation analysis

到稿日期:2019-03-18 返修日期:2019-06-30 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:西安市科学技术局科技创新引导项目(201805045YD23CG29(5));长安大学中央高校基本科研业务费专项资金(300102329203);长安大学研究生科研创新实践项目(300103002075)
This work was supported by the Science and Technology Innovation Guidance Project of Xi'an Science and Technology Bureau (201805045YD23CG29(5)), Fundamental Research Funds for the Central Universities, Chang'an University (CHD) (300102329203), postgraduate research innovation practice project of Chang'an University (300103002075).
通信作者:李刚(15229296166@chd.edu.cn)

1　引言

近年来，许多研究人员将典型关联分析模型的各种扩展形式应用于全基因组关联分析[1]并获得了极大的进展，这些扩展形式主要包括单核苷酸多态性数据（single Nucleotide Polymorphisms，SNPs）与基因数据[2]，以及 DNA 拷贝数变化与功能磁共振成像数据（functional Magnetic Resonance Imaging，fMRI）[3-4]之间的关联分析。本文首先利用存在于 116 个脑区并包含 183 个样本的 fMRI 数据集和具有特定基因的 SNPs 数据集[5-6]构造了两类模态数据，然后将典型关联分析（Canonical Correlation Analysis，CCA）模型及其扩展模型应用于两类模态数据，最后分析一种模态数据与另一种模态数据之间的最大关联性。目前，已有许多学者[2,5-9]开发出了很多 CCA 模型的不同形式来应对大规模研究中的挑战。

为了本研究的顺利展开，首先需要回顾该研究所涉及的统计学习模型。CCA 模型[5]是一种具有线性投影性质的设计方法，能够发现两组典型变量之间的最大关联性并降低特征向量维数；而稀疏典型关联分析模型（sparse Canonical Correlation Analysis，sCCA）[2,8]和联合稀疏典型关联分析模型（joint sparse Canonical Correlation Analysis，joint-sCCA）[9]是 CCA 模型的扩展形式，具有特殊的正则化约束，能够使这种惩罚模型更易于解释。CCA 模型本身具有局限性，仅能检测两组典型变量之间简单的线性关系，不能检测复杂的非线性关系。对此，Andrew 等提出了一种可以找到多模态数据间复杂非线性关系的深度典型关联分析（Deep Canonical Correlation Analysis，DCCA）模型[10]，而 Wang 等之后改进的深度典型相关自编码器模型（Deep Canonical Correlated Auto-Encoders，DCCAE）可以优化一般自编码器重构误差与神经网络节点特征输出之间的非线性相关组合[11]。

2　数学模型

2.1　典型关联分析

Hotelling 等提出的典型关联分析模型，通常被用于确定两组数据[7]之间的线性相关性。该模型为多模态数据研究协方差问题提供了一种思路，换言之，该模型只寻找与两类模态数据相关性最大的线性投影。

实验中，两类模态数据包含相同数量的样本和不同数量的高维特征，$\boldsymbol{X}\in R^{n\times p_x}$ 和 $\boldsymbol{Y}\in R^{n\times p_y}$ 是两类模态数据的定义形式，而 CCA 模型为了最大化关联性，试图找到两个线性投影向量 $\boldsymbol{u}\in R^{p_x}$ 和 $\boldsymbol{v}\in R^{p_y}$，如式（1）所示：

$$\begin{aligned}&\max_{\boldsymbol{u},\boldsymbol{v}}\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}\\&\text{s.t.}\ \ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}=1\end{aligned}\tag{1}$$

其中，数据矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 假设已经标准化，标准差为 1，均值为 0；$\boldsymbol{Xu}$ 和 $\boldsymbol{Yv}$ 为 CCA 模型的典型变量。

2.2　稀疏典型关联分析

CCA 被广泛用于线性关联性分析，但所使用的多模态数据集的高维特征数量远多于实际样本的数量。为此，许多实验都引入了带有不同惩罚项的 CCA 模型[2-3,12-13]，这种方式在一定程度上解决了使用特征维度高但样本数量较少的数据集时所面临的问题。它们对权重向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 执行不同的稀疏惩罚函数，以获得一些有研究价值的稀疏权重向量。稀疏 CCA[8]模型由 Witten 等提出，式（2）给出了该模型的函数方程，其对应的图解如图 1 所示。

$$\begin{aligned}&\max_{\boldsymbol{u},\boldsymbol{v}}\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}\\&\text{s.t.}\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}\leqslant 1,\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}\leqslant 1,\\&\quad P(\boldsymbol{u})\leqslant c_1,P(\boldsymbol{v})\leqslant c_2\end{aligned}\tag{2}$$

其中，P 是凸优化函数（如 L_1 范数），而 c_1 和 c_2 是决定 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 稀疏程度的权衡参数。CCA 模型的优化问题本可以利用 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ 和 $(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})^{-1}$ 解决，但高维数据矩阵可能是奇异的，因此可能会导致 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 和 $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$ 的逆矩阵不存在。假设两个数据矩阵中的特征是不相关的，则式（1）中的约束条件 $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}=1$ 变成了式（2）中的约束条件 $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}\leqslant 1,\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}\leqslant 1$，也正是式（1）中的约束条件使式（2）中的惩罚项可以最大化协方差 $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}$，而不是最大化 Pearson 相关系数 $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}/\boldsymbol{X}\boldsymbol{u}\boldsymbol{Y}\boldsymbol{v}$。这种方法是可取的，因为它能很大程度上降低计算 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ 和 $(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})^{-1}$ 的成本。

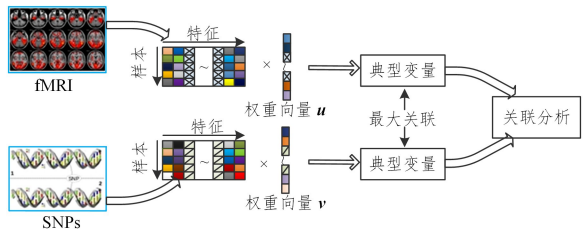


图 1　稀疏典型关联分析模型的图解

Fig. 1　Correlation analysis by sparse canonical correlation analysis algorithm

在该模型中，两类模态数据包括相同的样本容量 n 和不同的特征数 p 和 q，通过使用惩罚项来降低高维数据的维度。我们使用权重向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 来获取相应的典型变量，并进行关联分析来得到典型变量之间的最大关联性。

2.3　深度典型关联分析

随着深度神经网络在癌症检测、图像处理等领域的突破性应用，其使用率呈爆炸式增长，它可以表示为具有两层或多层结构的高维非线性函数。各种数据资源的积累和计算能力的提高促进了深度网络模型的使用；同时，随着数据集大小和复杂度的逐渐增加，深度网络模型的性能也得到了极大的优化。为了弥补前面提到的 CCA 模型和 sCCA 模型存在的不足，Andrew 等提出了具有数据预处理能力的 DCCA 模型来捕获多模态数据间的非线性映射。DCCA 模型如式（3）所示：

$$\begin{aligned}&\max_{\boldsymbol{W}_f,\boldsymbol{W}_g,\boldsymbol{U},\boldsymbol{V}}\frac{1}{N}Trace(\boldsymbol{U}^{\mathrm{T}}f(\boldsymbol{X})^{\mathrm{T}}g(\boldsymbol{Y})\boldsymbol{V})\\&\text{s.t.}\ \boldsymbol{U}^{\mathrm{T}}(\frac{1}{N}f(\boldsymbol{X})^{\mathrm{T}}f(\boldsymbol{X}))\boldsymbol{U}=\mathrm{I},\\&\quad\boldsymbol{V}^{\mathrm{T}}(\frac{1}{N}g(\boldsymbol{Y})^{\mathrm{T}}g(\boldsymbol{Y}))\boldsymbol{U}=\mathrm{I}\end{aligned}\tag{3}$$

其中，f 和 g 是两个深度神经网络，用于提高捕获线性和复杂非线性关系的能力；$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 是两个投影矩阵；$f(\boldsymbol{X})$ 和 $g(\boldsymbol{Y})$ 是这两个深度神经网络最后一层的输出。DCCA 目标函数利用 whitening 约束条件将所有预处理的样本数据连接在一起，因

此随机梯度下降算法在正常情况下并不适用。而 Wang 等提出的方法是使用大量的小批次数据来处理优化随机梯度[14]，这样便足以保证此模型的运行。

输入层的数据是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，而两个模态数据实际的节点数由真实数据的高维特征决定。输出层的数据是 $f(\boldsymbol{X})$ 和 $g(\boldsymbol{Y})$，其深度网络结构由 f 和 g 表示，这两个深度网络均包括两个分别用于处理线性不可分问题的隐藏层。

2.4 深度典型相关自编码器

深度典型相关自编码器[11]是一种由几个自编码器组成的改进 DCCA 模型，该模型优化了一般自动编码器重构误差与神经网络学习表示之间的非线性关联组合。DCCAE 模型的目标函数如式(4)所示：

$$\max_{W_f,W_g,W_p,W_q,U,V}\frac{1}{N}Trace(\boldsymbol{U}^{\mathrm{T}}f(\boldsymbol{X})^{\mathrm{T}}g(\boldsymbol{Y})\boldsymbol{V})-$$

$$\frac{\gamma}{N}\sum_{i=1}^{N}(\|x_i-p(f(x_i))\|^2-\|y_i-q(g(y_i))\|^2) \tag{4}$$

s. t. 相应约束条件与式(3)一致

其中，γ 是权衡参数。该函数是为了找到两个正则化的深度神经网络表示形式 $f(\boldsymbol{X})$ 和 $g(\boldsymbol{Y})$，以及两个重构网络表示形式 $p(f(x_i))$ 和 $q(g(y_i))$。DCCAE 模型是受 DCCA 和最小重构特征误差的启发而获得多模态数据的最大关联性的一种有效模型，其包括两个特征提取网络 f 和 g，以及两个重建网络 p 和 q。该模型采用随机梯度法对目标函数进行优化，有效地保证了多模态数据相关性的最佳效果。与 CCA 和 DCCA 相比，DCCAE 模型可以进一步提取特征，使数据拟合效果更好。

2.5 深度主成分相关自编码器

在 DCCAE 和主成分分析(Principal Component Analysis，PCA)方法的启发下，本研究设计了一种新颖的模型——深度主成分相关自编码器(Deep Principal Correlated Auto-Encoders，DPCAE)，它由两个反向传播神经网络和两个由限制玻尔兹曼机(Restricted Boltzmann Machine，RBM)组成的深度信念网络(Deep Belief Network，DBN)构成，同时对模型的最大关联性进行积分。

值得注意的是，该模型使用两个深度信念网络 f 和 g 来提取相应隐藏层单元的特征表示，用两个反向传播神经网络 p 和 q 来提取顶层单元的特征表示，并分别在每种模态下对整个深度信念网络进行微调。用容量分别为 $n\times p_x$ 和 $n\times p_y$ 的 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 定义两种模态数据，两种模态数据具有相同的样本数 n，p_x 和 p_y 分别是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的特征维数，F_k 和 F_h 是 PCA 方法的表示形式，p_{x_1} 和 p_{y_1} 是经过线性降维后的特征维数，p_{x_2} 和 p_{y_2} 是经过 DBN 网络训练之后的特征维数，而 $p_{x_{3i}}$ 和 $p_{y_{3i}}$ $(i=1,2,\cdots,m)$ 是使用 RBM 重构特征之前输入层的特征维数。DPCAE 的目标函数如式(5)所示，其相应模型的示意图如图 2 所示。

$$\max_{\theta}\frac{1}{N}\mathrm{Trace}(\boldsymbol{U}^{\mathrm{T}}f(F_k(\boldsymbol{X}))^{\mathrm{T}}g(F_h(\boldsymbol{Y}))\boldsymbol{V})-\frac{\gamma_1}{N}\sum_{i=1}^{m}\sum_{j=1}^{N}$$

$$[(\|F_k(x_{ij})-\theta_1 r_{1i}(F_k(x_{ij}))\|^2)+(\|F_h(y_{ij})-\theta_2 r_{2i}$$

$$(F_h(y_{ij}))\|^2)]-\frac{\gamma_2}{N}[(\|L-\boldsymbol{W}_p p(f(F_k(X)))\|^2)+$$

$$(\|L-\boldsymbol{W}_q q(g(F_h(\boldsymbol{Y})))\|^2)]-P(\boldsymbol{W}^*) \tag{5}$$

$$\text{s. t. } \boldsymbol{U}^{\mathrm{T}}(\frac{1}{N}f(F_k(\boldsymbol{X}))f(F_k(\boldsymbol{X}))^{\mathrm{T}}+r_1\boldsymbol{I})\boldsymbol{U}=\boldsymbol{I},$$

$$\boldsymbol{V}^{\mathrm{T}}(\frac{1}{N}g(F_h(\boldsymbol{Y}))g(F_h(\boldsymbol{Y}))^{\mathrm{T}}+r_2\boldsymbol{I})\boldsymbol{U}=\boldsymbol{I},$$

$$u_i^{\mathrm{T}}f(F_k(\boldsymbol{X}))^{\mathrm{T}}g(F_h(\boldsymbol{Y}))v_j=0,\ \text{for } i\neq j$$

其中，$P(\boldsymbol{W}^*)$ 是凸优化函数 $P(\boldsymbol{W}_p)$ 和 $P(\boldsymbol{W}_q)$ 的联合函数，而这些凸优化函数的形式是可以防止模型过拟合的岭函数，该函数形式在一定程度上保证了模型的非奇异性。$P(\boldsymbol{W}^*)$ 是平方 L_2 范数在 $\boldsymbol{W}_p$ 和 $\boldsymbol{W}_q$ 上的和，即 $P(\boldsymbol{W}^*)=\mu_p\|\boldsymbol{W}_p\|_2^2+\mu_q\|\boldsymbol{W}_q\|_2^2$。$\theta$ 是 $\boldsymbol{W}_p$，$\boldsymbol{W}_q$，θ_1，θ_2，$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的组合，μ_p 和 μ_q 是调整两类函数合理性的权值衰减参数，γ_1 和 γ_2 是调整参数。

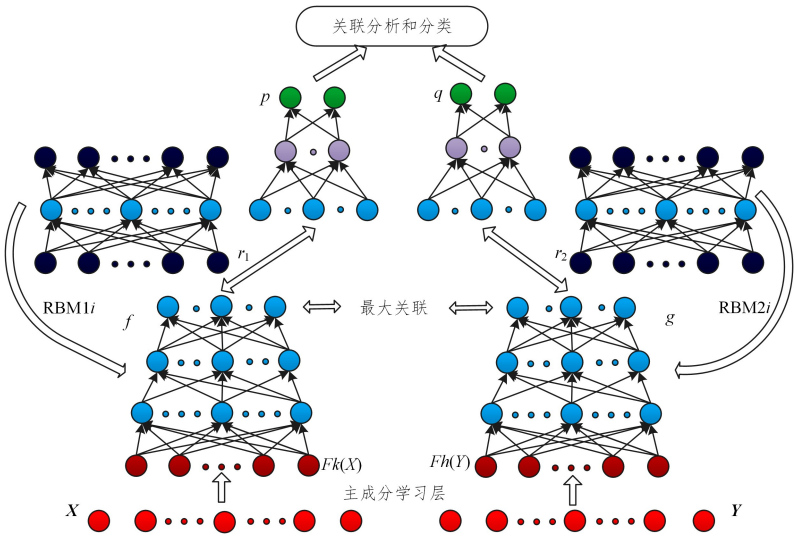


图 2 DPCAE 模型的示意图

Fig. 2 Schematic diagram of DPCAE

当计算 $(f(F_k(\boldsymbol{X}))f(F_k(\boldsymbol{X}))^{\mathrm{T}})^{-1}$ 和 $(g(F_h(\boldsymbol{Y}))g(F_h(\boldsymbol{Y}))^{\mathrm{T}})^{-1}$ 的算式时，$f(F_k(\boldsymbol{X}))f(F_k(\boldsymbol{X}))^{\mathrm{T}}$ 和 $g(F_h(\boldsymbol{Y}))g(F_h(\boldsymbol{Y}))^{\mathrm{T}}$ 可能是奇异的。但上述两种算式的计算结果又极其重要，因此矩阵正则化形式通常按照式(6)和式(7)进行定义，

以确保其非奇异性。

$$\frac{1}{N}f(F_k(\boldsymbol{X}))f(F_k(\boldsymbol{X}))^{\mathrm{T}\sim}=\frac{1}{N}f(F_k(\boldsymbol{X}))f(F_k(\boldsymbol{X}))^{\mathrm{T}}+\delta_1\boldsymbol{I} \tag{6}$$

$$\frac{1}{N}g(F_h(\boldsymbol{Y}))f(F_k(\boldsymbol{Y}))^{\mathrm{T}\sim}=\frac{1}{N}\mathrm{g}(\mathrm{F_h}(\boldsymbol{Y}))g(F_h(\boldsymbol{Y}))^{\mathrm{T}}+\delta_2\boldsymbol{I} \tag{7}$$

其中，$\delta_1,\delta_2>0$ 是方差多变量分析的正则化参数。

DPCAE 模型由两个 BP 网络和两个由 RBM 构成的深度信念网络组成。每个深度信念网络的顶层相关系数最高。利用最大关联系数对每个 BP 网络的输入层进行优化，从而得到多模态影像遗传数据的最佳关联分析和分类结果。式(5)的方程用式(8)代替：

$$\begin{aligned}Z(R_1,R_2,P,Q)=\max_{\theta}\|\sum\nolimits_{11}^{-1/2}\sum\nolimits_{12}\sum\nolimits_{22}^{-1/2}\|-\|F_k(\boldsymbol{X})-R_1(R_1'R_1)^{-1}R_1'F_k(\boldsymbol{X})\|_2^2-\|F_h(\boldsymbol{Y})-R_2(R_2'R_2)^{-1}R_2'F_h(\boldsymbol{Y})\|_2^2-\|L-P(P'P)^{-1}P'L\|_2^2-\|L-Q(Q'Q)^{-1}Q'L\|_2^2\end{aligned} \tag{8}$$

其中，$R_1=r_1(F_k(\boldsymbol{X}))\in R^{n\times p_{x_3}}$，$R_2=r_2(F_h(\boldsymbol{Y}))\in R^{n\times p_{y_3}}$，$P=p(f(F_k(\boldsymbol{X})))\in R^{n\times p_{x_2}}$，$Q=q(g(F_h(\boldsymbol{Y})))\in R^{n\times p_{y_2}}$，$\sum_{11}=f(F_k(\boldsymbol{X}))f(F_k(\boldsymbol{X}))$，$\sum_{12}=f(F_k(\boldsymbol{X}))g(F_h(\boldsymbol{Y}))$ 和 $\sum_{22}=g(F_h(\boldsymbol{Y}))g(F_h(\boldsymbol{Y}))$。

线性 CCA 和线性 sCCA 模型利用最优投影向量 $\boldsymbol{U}(\boldsymbol{u})$ 和 $\boldsymbol{V}(\boldsymbol{v})$ 寻求最大关联性，两者对应的函数已在式(2)和式(3)中给出。本文之所以回顾 DCCA 和 DCCAE，主要是为了进一步阐述 CCA 不能分析典型变量间复杂非线性关系的局限性。与一般的 DNN 模型相比，DPCAE 模型结合多个 RBM 和线性降维的方法，搭建了更有效的深度信念网络，不仅加快了机器学习的速度，而且抛弃了含有较少信息的维度；此外，不同于非线性的 DCCAE 模型，DPCAE 模型利用深层信念网络调整网络的层内参数，并利用 BP 网络调整顶层网络参数，以进一步自上而下地微调整个网络层的参数，使得在多模态数据上应用 PCA 进行数据拟合时获得了较为理想的非线性关联效果。特别地，DPCAE 模型还进一步优化了重要的特征信息和参数信息，这在一定程度上很好地表示了多模态数据之间的关联性。

为了获得更高的计算效率，本文应用一阶导数信息和共轭梯度下降法[15]求解优化问题。在每次迭代过程中，由 RBM 和 BP 算法组成的深度置信网络都会将共轭梯度反向传递给网络的每一层；此外，该模型还采用正则化方法以避免过拟合。使用共轭梯度下降和 BP 时必须计算式(5)的梯度，其梯度公式如下：

$$\begin{aligned}\frac{\partial Z(R_1,R_2,P,Q)}{\partial R_1}=R_1\sum\nolimits_{11}^{-1/2}U_1D_1U_1'\sum\nolimits_{11}^{-1/2}+R_2\sum\nolimits_{22}^{-1/2}V_1U_1'\sum\nolimits_{22}^{-1/2}+2F_k(\boldsymbol{X})F_k(\boldsymbol{X})'R_1(R_1'R_1)^{-1}-2R_1(R_1'R_1)^{-1}R_1'F_k(\boldsymbol{X})F_k(\boldsymbol{X})'R_1(R_1'R_1)^{-1}\end{aligned} \tag{9}$$

$$\begin{aligned}\frac{\partial Z(R_1,R_2,P,Q)}{\partial R_2}=R_2\sum\nolimits_{22}^{-1/2}V_1D_1V_1'\sum\nolimits_{22}^{-1/2}+R_1\sum\nolimits_{11}^{-1/2}U_1V_1'\sum\nolimits_{22}^{-1/2}+2F_h(\boldsymbol{Y})F_h(\boldsymbol{Y})'R_2(R_2'R_2)^{-1}-2R_2(R_2'R_2)^{-1}R_2'F_h(\boldsymbol{Y})F_h(\boldsymbol{Y})'R_2(R_2'R_2)^{-1}\end{aligned} \tag{10}$$

$$\frac{\partial Z(R_1,R_2,P,Q)}{\partial P}=2LL'P(P'P)^{-1}-2P(P'P)^{-1}P'LL'P(P'P)^{-1} \tag{11}$$

$$\frac{\partial Z(R_1,R_2,P,Q)}{\partial Q}=2LL'Q(Q'Q)^{-1}-2Q(Q'Q)^{-1}Q'LL'Q(Q'Q)^{-1} \tag{12}$$

DPCAE 算法的具体流程如算法 1 所示。

算法 1　DPCAE 算法

输入：$\boldsymbol{X}\in R^{n\times p_x}$，$\boldsymbol{Y}\in R^{n\times p_y}$，表型 $\boldsymbol{L}\in R^{n\times 1}$，初始化网络 f_1，g_1，p_1，q_1，r_{11}，r_{21}

输出：使用 $\hat{f}(F_k(\boldsymbol{X}))\in R^{n\times p_{x_1}}$ 和 $\hat{g}(F_h(\boldsymbol{Y}))\in R^{n\times p_{y_1}}$ 优化的网络 $\hat{f}_1$ 和 $\hat{g}_1$，使用 $\hat{p}(\hat{f}(F_k(\boldsymbol{X})))\in R^{n\times p_{x_2}}$ 和 $\hat{q}(\hat{g}(F_h(\boldsymbol{Y})))\in R^{n\times p_{y_2}}$ 优化的网络 $\hat{p}_1$ 和 $\hat{q}_1$，使用 $\hat{r}_1(\hat{f}_i(F_k(\boldsymbol{X})))\in R^{n\times p_{x_{3i}}}$ 和 $\hat{r}_2(\hat{g}_i(F_h(\boldsymbol{Y})))\in R^{n\times p_{y_{3i}}}$ 优化的网络 $\hat{r}_{11}$ 和 $\hat{r}_{21}$；

1. $\hat{f}(F_k(\boldsymbol{X}))\leftarrow\hat{f}_1(F_k(\boldsymbol{X}))$，$\hat{g}(F_h(\boldsymbol{Y}))\leftarrow\hat{g}_1(F_h(\boldsymbol{Y}))$，$\hat{p}(\hat{f}(F_k(\boldsymbol{X})))\leftarrow\hat{p}_1(\hat{f}(F_k(\boldsymbol{X})))$，$\hat{q}(\hat{g}(F_h(\boldsymbol{Y})))\leftarrow\hat{q}_1(\hat{g}(F_h(\boldsymbol{Y})))$，$\hat{r}_1(\hat{f}_i(F_k(\boldsymbol{X})))\leftarrow\hat{r}_{11}(\hat{f}_i(F_k(\boldsymbol{X})))$，$\hat{r}_2(\hat{g}_i(F_h(\boldsymbol{Y})))\leftarrow\hat{r}_{21}(\hat{g}_i(F_h(\boldsymbol{Y})))$，$i=1,2,\cdots,m$；
2. 迭代次数 $k\leftarrow0$；
3. while $k<$最大迭代次数 and 不收敛 do
4. $\hat{p}_1\leftarrow$式(9)和式(10)；
5. $\hat{f}\xleftarrow{\text{堆栈}}\hat{r}_{11}\leftarrow\mathrm{RBM}_{1i}(\hat{r}_{11},|\nabla Z(R_1,R_2,P,Q)|_{R_1=\hat{r}_{11},R_2=\hat{r}_{21}})$；
6. $\hat{g}\xleftarrow{\text{堆栈}}\hat{r}_{21}\leftarrow\mathrm{RBM}_{2i}(\hat{r}_{21},\nabla Z(R_1,R_2,P,Q)|_{R_1=\hat{r}_{11},R_2=\hat{r}_{21}})$；
7. $\nabla Z(R_1,R_2,P,Q)|_{P=\hat{p}_1,Q=\hat{q}_1}\leftarrow$式(11)和式(12)；
8. $\hat{p}_1\leftarrow$反向传播$(\hat{p}_1,\nabla Z(R_1,R_2,P,Q)|_{P=\hat{p}_1,Q=\hat{q}_1})$；
9. $\hat{q}_1\leftarrow$反向传播$(\hat{q}_1,\nabla Z(R_1,R_2,P,Q)|_{P=\hat{p}_1,Q=\hat{q}_1})$；
10. $\hat{p}\leftarrow$前向传播$(\hat{p}_1(\hat{f}(F_k(X))))$；
11. $\hat{q}\leftarrow$前向传播$(\hat{q}_1(\hat{g}(F_h(Y))))$；
12. $k\leftarrow k+1$
13. return $\hat{f},\hat{g},\hat{p}_1,\hat{q}_1,\hat{p},\hat{q}$。

3　实验分析

针对多模态数据集的实验，DPCAE 模型的工作主要集中在两个方面：分类(使用 SNPs 位点数据和 fMRI 成像数据对不同的受试者进行分类)和关联分析(探索 SNPs 位点数据与 fMRI 成像数据之间的关联性)。分类实验验证了 DPCAE 算法的分类能力，关联分析实验验证了 DPCAE 算法在寻找最大关联性方面的性能。使用共轭梯度下降法选择所有超参数，包括能量函数、激活函数类型、正则化参数、学习率、权衡参数、最大迭代次数、深度网络类型、层数、不同网络层节点数等[16]。实验环境由 8 台联想服务器进行配置，每台服务器均搭载 2 个四核八线程的 Intel (R) Core (TM) i7-7400 CPU，128GB 内存，8 台服务器构成了一个 Spark 高可用分布式集

群[17]，其中安装了 2.6.31-16-generic 内核和 64 位 Ubuntu 操作系统，以及 java1.8.0_191-b12，python3.7，Hadoop2.9.1 和 Spark2.3.0。实验时，使用 Python 语言在 Spark 分布式集群上进行调试和结果分析。

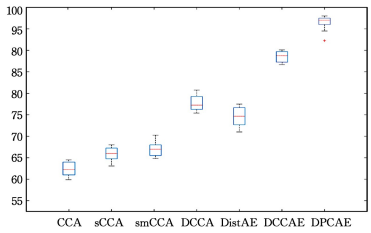
3.1 分类精度的比较

在 fMRI 成像数据分类和 SNPs 位点数据分类方面，DPCAE 模型首先使用两类模态数据进行训练，然后利用训练后的网络进行分类，最后全面分析并比较了该模型与其他先进算法(CCA，sparse CCA，smCCA，DCCA，distAE，DCCAE)的性能差异。对于多模态影像数据，通过将已训练的模态数据与实际数据进行比较来评价其分类精度。该分类方法通过箱型图可观测各模型对临床数据分析的精度，其分类精度的计算公式如式(13)所示：

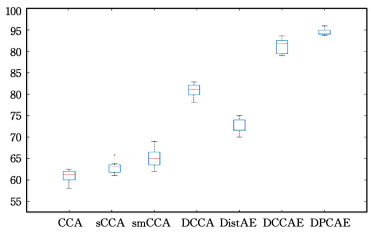
$$Pre=TP/(TP+FP) \tag{13}$$

其中，TP 为经各模型训练之后的模态数据被正确分类为正样本的个数，FP 为经各模型训练之后的模态数据被错误分类为正样本的个数。

一些有研究价值的预处理方法也被应用在多模态数据[14]上，如数据增长、数据标准化等。与至少需要两类模态数据作为输入的 CCA 模型类似，该实验使用了 fMRI 成像数据和 SNPs 位点数据的组合形式；在模态数据组合的基础上，对每个模型的性能都进行实验，结果如图 3 所示。图 3(a)描述了使用 fMRI 成像数据对不同受试者的分类能力，图 3(b)描述了使用 SNPs 位点数据对不同受试者的分类能力，其中，X 轴表示几种不同的实验模型，Y 轴表示分类精度(单位为%)。



(a)分类 fMRI



(b)分类 SNPs

图 3 各模型对两种模态数据分类能力的对比

Fig. 3 Comparison of power of each model on classifying two types of modalities (sub-figs a-b)

从图 3 中可以看出，在使用两类模态数据的组合形式对不同的受试者进行分类时，与其他基于 CCA 的模型和基于 DNN 的模型相比，DPCAE 模型具有更好的分类精度。在分类精度上，基于 DNN 的模型(DCCA 和 DCCAE)优于所有基于 CCA 的模型和 distAE 模型，但却差于 DPCAE 模型，这种效果上的显著差异是由于加入了 DNN 输出和正则化参数来评估样本方差造成的。然而，distAE 模型的分类性能比 DCCA 和 DCCAE 的性能差，这是由于投影样本对之间的平均差较弱所致。显而易见，DPCAE 模型在两类模态数据上的分类精度均较好，其中 SNPs 位点数据的分类精度超过了 90%，fMRI 成像数据的分类精度也超过了 95%。本实验还有一点须指出，即 fMRI 成像数据的分类精度高于 SNPs 位点数据的分类精度，这可能是由于 SNPs 位点数据因样本或环境因素导致连锁不平衡状态不准确，从而删除某些重要的基因片段所致。探索影像基因组学的不同原始特征如何影响 SNPs 位点数据和 fMRI 成像数据的分类精度是一个很有意义的研究课题，然而 DPCAE 模型的数据表示是由不同的深度网络构成的，每个隐含层都使用非线性激活函数，分析每个原始典型变量在深度网络表示中是如何显示的并不容易，因此针对每个原始生物标志物变量的识别也是一个挑战。

3.2 聚类性能的评估

归一化互信息是一种信息度量形式，可以将其看作一种典型随机变量包含另一种典型随机变量的信息量。分类错误率是分类错误的样本数占总样本数的比例，其结果是为了进一步证明分类精度的有效性。为了本节实验的顺利进行，先给出归一化互信息的计算公式：

$$\mathrm{Min}(\boldsymbol{X}_{\mathrm{tr}};\boldsymbol{Y}_{\mathrm{tr}})=2I(\boldsymbol{X}_{\mathrm{tr}};\boldsymbol{Y}_{\mathrm{tr}})/H(\boldsymbol{X}_{\mathrm{tr}})+H(\boldsymbol{Y}_{\mathrm{tr}}) \tag{14}$$

$$\text{s.t. } I(X_{\mathrm{tr}};Y_{\mathrm{tr}})=\sum_{\boldsymbol{x}_{\mathrm{tr}}}\sum_{\boldsymbol{y}_{\mathrm{tr}}}p(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{y}_{\mathrm{tr}})*\log\{p(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{y}_{\mathrm{tr}})/[p(\boldsymbol{x}_{\mathrm{tr}})p(\boldsymbol{y}_{\mathrm{tr}})]\}$$

$$H(\boldsymbol{X}_{\mathrm{tr}})=-\sum_i p(\boldsymbol{x}_{\mathrm{tri}})\log p(\boldsymbol{x}_{\mathrm{tri}})$$

$$H(\boldsymbol{Y}_{\mathrm{tr}})=-\sum_i p(\boldsymbol{y}_{\mathrm{tri}})\log p(\boldsymbol{y}_{\mathrm{tri}})$$

其中，$p(\boldsymbol{x}_{tr},\boldsymbol{y}_{tr})$是两个经训练后的两组数据 $\boldsymbol{X}_{tr}$ 和 $\boldsymbol{Y}_{tr}$ 的联合分布，$p(\boldsymbol{x}_{tr})$和 $p(\boldsymbol{y}_{tr})$分别是 $\boldsymbol{x}_{tr}$ 和 $\boldsymbol{y}_{tr}$ 的概率分布，$H(\boldsymbol{X}_{tr})$和 $H(\boldsymbol{Y}_{tr})$分别是 $\boldsymbol{X}_{tr}$ 和 $\boldsymbol{Y}_{tr}$ 的信息熵，而 $I(\boldsymbol{X}_{tr};\boldsymbol{Y}_{tr})$是联合分布 $p(\boldsymbol{x}_{tr},\boldsymbol{y}_{tr})$与乘积分布 $p(\boldsymbol{x}_{tr})p(\boldsymbol{y}_{tr})$的相对熵。

根据不同模型训练的数据，对两类模态数据的分类错误率进行计算：

$$Err=(FP+FN)/(P+N) \tag{15}$$

其中，FP 是经训练的两类模态数据实际为负样本但被分类器分为正样本的个数，FN 是两类模态数据实际为正样本但被分类器分为负样本的个数，而 $P+N$ 是两类不同模态数据的总样本数。

本实验首先将已投影的 SNPs 位点数据和 fMRI 成像数据输入到两个聚类中，然后根据样本标签分别对聚类的性能进行评估，最后估计所提取特征空间中的类距离。为了解释潜在的非凸聚类形状，该实验还应用了谱聚类算法[18]。另外，已知实验中使用的 k 最邻近算法能以二进制加权方式构造两类模态映射样本数据，那么这些映射样本将采用归一化的 Euclidean 特征向量图嵌入到 R^2 中，最后运行嵌入后的 k-means 程序，并找到一个可用的分割样本。k-means 程序用关联性最高时的特征提取网络节点输出作为聚类初始化运行了 5 次，并将最佳的 k-means 值作为基础解释谱聚类算法。本实验还利用归一化互信息[19]估计了聚类能力，并以调优集

合{3,5,10,20,30}为最邻近 k 的大小。实验结果如表 1 所列。

表 1　各经典研究方法下两种模态数据的性能评估

Table 1　Performance evaluation of each typical research methods on dataset of SNPs and fMRI

（单位：%）

模型	S-归一化	S-错误率	F-归一化	F-错误率
CCA	49.1	22.5	51.6	20.8
sCCA	57.6	19.2	59.8	17.5
smCCA	59.5	18.3	64.2	16.3
DCCA	86.3	8.1	86.9	5.8
distAE	65.7	12.6	73.4	10.9
DCCAE	89.6	5.8	92.3	4.6
DPCAE	93.4	4.3	94.1	3.3

由表 1 可知，DPCAE 模型对不同对象进行聚类时所获得的归一化互信息（包括 SNPs 位点数据和 fMRI 成像数据的归一化互信息，将其简写为 S-归一化和 F-归一化）均优于基于 CCA 和其他基于 DNN 的模型。SNPs 位点数据和 fMRI 成像数据的归一化互信息分别为 93.4%和 94.1%，这说明不同研究对象的归一化互信息差异较大。虽然 distAE 模型的预测结果显示，该模型确实能对一些集群执行更好的分割，但其比其他基于 DNN 的模型都差。另一方面，基于 CCA 的模型可以逼近同类变量，但是对非同类变量的聚类效果却很不理想，这主要是因为原始数据太复杂，不能很好地应用线性映射关系对其聚类。总的来说，DPCAE 显示了对嵌入后的 SNPs 位点数据和 fMRI 成像数据进行聚类时归一化互信息的最佳结果。该实验希望可以用一个简单的线性分类器在 DPCAE 模型上获得很好的分类精度，因此利用调优集合选择 SVM 超参数[20]，进而在投影数据集上对线性 SVM 进行训练（现在使用的是两类模态数据的样本标签）。各模型的分类错误率（包括 SNPs 位点数据和 fMRI 成像数据的分类错误率，将其简写为 S-错误率和 F-错误率）如表 1 所列，可以看出这些分类错误率与聚类结果一致。由于该算法在高维特征表示上用一个普通的线性分类器替代了复杂的非线性分类器分类数据，因此使得在数据分类过程中相应程序的复杂度大大降低，有效地对低维投影进行了训练并获得了最小的分类错误率。

3.3　最大关联性分析

为了能够更好地比较 DPCAE 模型与其他先进模型之间的性能差异，本实验还将两类模态数据组合起来作为原始数据，并将其作为顶层节点输入到 DPCAE 网络模型中，之后在顶层输出的不同维度下比较了 DPCAE 模型与其他先进模型的最大关联性。这些模型都进行了 40 次最大关联性的实验，每次使用 50%的多模态数据进行训练，30%的数据用于超参数调整，剩余 20%的数据用于最终测试。

由式(5)可知，有 6 个最优化的参数被用来计算不同维度下顶层输出节点的最大关联系数，将这些先进模型分别做最大关联性分析：

$$\sigma = Corr(\boldsymbol{X}_{\text{train}}\boldsymbol{u}_{\text{test}}, \boldsymbol{Y}_{\text{train}}\boldsymbol{v}_{\text{test}}) \tag{16}$$

其中，$\boldsymbol{X}_{\text{train}}$ 和 $\boldsymbol{Y}_{\text{train}}$ 为经过训练的输出数据集，$\boldsymbol{u}_{\text{test}}$ 和 $\boldsymbol{v}_{\text{test}}$ 为用于最终测试的最优典型变量。将这些模型分别进行多次实验，选择 $\arg\max\sigma$ 作为最大关联系数值。

基于 DNN 的模型与基于 CCA 的模型在顶层输出的不同维度上的最大关联性比较如图 4 所示。

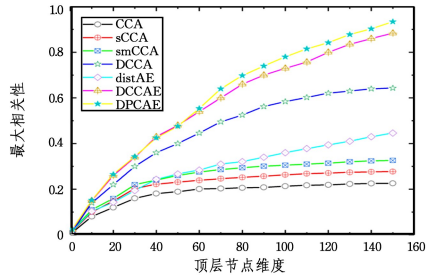


图 4　各模型在顶层输出维度不同时的最大关联性比较

Fig. 4　Comparison of highest correlation of CCA-based models and DNN-based models on different number of dimensions

图 4 显示了 CCA 模型、具有 L_1 惩罚项的 sCCA 模型、smCCA 模型、具有非线性映射关系的 DCCA 模型、distAE 模型、DCCAE 模型以及 DPCAE 模型在顶层输出维度变化时所获得的不同最大关联性。DCCAE 模型倾向于在一个小范围内发现微弱的非线性关系；而 DPCAE 模型能够在较大范围内发现典型变量间的非线性关系，其最大关联分析的能力远超 DCCAE 模型。特别需要注意的是，DPCAE 模型中应用的 PCA 技术可以很好地适应多模态数据；同样地，用于提取隐藏节点特征的两个深层网络和用于提取顶层特征并分别在每一类模态上调节整个深度信念网络的两个 BP 网络，均能提高模型的自学习自适应能力，以及增强多模态影像数据顶层节点的最大关联效果。从图中可以发现，DPCAE 模型确实能比基于 CCA 的模型和其他基于 DNN 的模型检测出更高的关联性。

结束语　本文对基于 CCA（CCA，sCCA 和 smCCA）和基于 DNN（DCCA，distAE，DCCAE 和 DPCAE）的模型进行了多次实验。结果表明，基于 DNN 的模型在各个方面的性能都优于基于 CCA 的模型，其中综合性能表现最好的是应用堆栈 RBM 的多层信念网络和 PCA 的 DPCAE 模型，因此基于 DNN 的模型在多模态数据关联性分析中可以获得更好的表现。DPCAE 模型不仅提供了一个灵活的非线性映射，而且提供了一个简单的线性映射，同时还具有不需要计算内积的优势。在实验研究的基础上，重新思考不同函数类型的基本性能和相应的惩罚项是很有吸引力的研究内容。一方面，基于自编码器的模型是建立在输出变量能够准确重构输入层变量的基础上的；另一方面，CCA 模型只局限性地关注了多模态学习特征如何预测其他特征，而忽略了重建多模态的能力。当两种模式不相关时，考虑到样本标签，CCA 方法有望取得较好的效果。各种算法中的惩罚项也有着至关重要的影响，客观地说，DPCAE 模型的惩罚条件较强，但还不够强，以后的实验中还需要一个更严格的惩罚条件来学习相互独立的多模态特征表示[21-23]。

参 考 文 献

[1]　NAYLOR M G，XIHONG L，WEISS S T，et al. Using Canonical Correlation Analysis to Discover Genetic Regulatory Variants [J]. Plos One，2010，5(5)：1-6.

[2]　PARKHOMENKO E，TRITCHLER D，BEYENE J. Sparse Ca-

nonical Correlation Analysis with Application to Genomic Data Integration [J]. Statistical Applications in Genetics and Molecular Biology, 2009, 8(1): 1-34.

[3] WAAIJENBORG S, VERSELEWEL D W H, PHILIP C, et al. Quantifying The Association between Gene Expressions and DNA-markers by Penalized Canonical Correlation Analysis [J]. Statistical Applications in Genetics & Molecular Biology, 2008, 7(1): 1-29.

[4] WITTEN D M, TIBSHIRANI R J. Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data [J]. Statistical Applications in Genetics & Molecular Biology, 2009, 8(1): 1-27.

[5] CAO S, QIN H, GOSSMANN A, et al. Unified Tests for Fine-scale Mapping and Identifying Sparse High-dimensional Sequence Associations [J]. Bioinformatics, 2016, 32(3): 330-337.

[6] DENG S P, HU W, CALHOUN V D, et al. Integrating Imaging Genomic Data in The Quest For Biomarkers for Schizophrenia Disease [J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2017, 15(5): 1480-1491.

[7] HOTELLING H. Relations Between Two Sets of Variates [J]. Biometrika, 1936, 28(3/4): 321-377.

[8] WITTEN D M, ROBERT T, TREVOR H. A Penalized Matrix Decomposition with Applications to Sparse Principal Components and Canonical Correlation Analysis [J]. Biostatistics, 2009, 10(3): 515-534.

[9] FANG J, LIN D, SCHULZ C, et al. Joint Sparse Canonical Correlation Analysis for Detecting Differential Imaging Genetics Modules [J]. Bioinformatics, 2011, 32(22): 3480-3488.

[10] ANDREW G, ARORA R, BILMES J, et al. Deep Canonical Correlation Analysis[C]// Proceedings of the International Conference on Machine Learning. 2013: 1247-1255.

[11] WANG W, ARORA R, LIVESCU K, et al. On Deep Multi-view Representation Learning[C]// Proceedings of International Conference on Machine Learning. 2015: 1083-1092.

[12] PARKHOMENKO E, TRITCHLER D, BEVENE J. Genome-wide Sparse Canonical Correlation of Gene Expression with Genotypes [J]. Bmc Proceedings, 2007, 1(9): 1-5.

[13] CAO K A L, MARTIN P G, ROBERT-GRANIE C, et al. Sparse Canonical Methods for Biological Data Integration: Application to a Cross-platform Study [J]. Bmc Bioinformatics, 2009, 10(1): 1-17.

[14] WANG W, ARORA R, LIVESCU K, et al. Unsupervised Learning of Acoustic Features via Deep Canonical Correlation Analysis[C]// Proceedings of IEEE International Conference on Acoustics. 2015: 1-5.

[15] DAI Y H, LIAO L Z, LI D. On Restart Procedures for The Conjugate Gradient Method [J]. Numerical Algorithms, 2004, 35(2/3/4): 249-260.

[16] HU W, CAI B, CALHOUN V, et al. Multi-modal Brain Connectivity Study Using Deep Collaborative Learning [J]. Springer Nature America, 2018, 7(4): 1-9.

[17] ZAHARIA M, CHOWDHURY M, FRANKLIN M J, et al. Spark: Cluster Computing with Working Sets [J]. HotCloud, 2010, 10(10): 95.

[18] NG A Y, JORDAN M I, WEISS Y. On Spectral Clustering: Analysis and an Algorithm[C]// Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. 2001: 1-8.

[19] CAI D, HE X, HAN J. Document Clustering Using Locality Preserving Indexing [J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(12): 1624-1637.

[20] GHADDAR B, NAOUMSAWAYA J. High Dimensional Data Classification and Feature Selection Using Support Vector Machines [J]. European Journal of Operational Research, 2018, 265(3): 86-93.

[21] SOHN K, SHANG W, LEE H. Improved Multimodal Deep Learning with Variation of Information[C]// Proceedings of the International Conference on Neural Information Processing Systems. 2014: 2141-2149.

[22] SRIVASTAVA N, SALAKHUTDINOV R. Multimodal Learning with Deep Boltzmann Machines [J]. Journal of Machine Learning Research, 2014, 15(8): 1-9.

[23] HU W, CAI B, ZHANG A, et al. Deep Collaborative Learning with Application to Multimodal Brain Development Study [J]. IEEE Transactions on Biomedical Engineering, 2019, 7(10): 1-8.

LI Gang, born in 1975, associate professor, postgraduate supervisor, is not member of China Computer Federation. His main research interests include image processing and pattern recognition, machine learning and multi-mode biomedical information fusion