

基于典型相关稀疏自编码器的精神分裂症的分类

李刚, 韩德鹏, 刘强伟, 王超, 李莹
长安大学电子与控制工程学院, 陕西 西安 710064

【摘要】通过结合大脑核磁共振成像和基因组信息进行全面系统的分析,影像遗传学已被广泛用于帮助诊断和治疗精神疾病(例如精神分裂症)。本文采用单核苷酸多态性数据和功能性磁共振成像数据联合分析,提出深度典型相关稀疏自编码器模型,探索两类数据之间的非线性关联并进行降维,对精神分裂症患者和健康对照进行分类。最后,实验结果表明,使用深度典型相关稀疏自编码器模型比其他传统模型具有更高的分类准确性。

【关键词】精神分裂症;单核苷酸多态性;功能性核磁共振成像;典型相关稀疏自编码器

【中图分类号】R318;R749.3

【文献标志码】A

【文章编号】1005-202X(2020)03-0391-06

Classification of schizophrenia based on deep canonically correlated sparse autoencoder

LI Gang, HAN Depeng, LIU Qiangwei, WANG Chao, LI Ying
School of Electronic Control, Chang'an University, Xi'an 710064, China

Abstract: Imaging genetics has been widely used to aid the diagnosis and treatment of mental illnesses (e.g. schizophrenia), by combining magnetic resonance imaging of the brain and genomic information for a comprehensive and systematic analysis. Herein the joint analysis of single nucleotide polymorphism data and functional magnetic resonance imaging data is carried out for the comprehensive study on schizophrenia. Deep canonically correlated sparse autoencoder is proposed to explore the nonlinear relationships between two types of data and reduce dimensions, and then classify schizophrenia patients from healthy controls. The experimental results reveal that compared with other conventional models, deep canonically correlated sparse autoencoder model can achieve a higher classification accuracy.

Keywords: schizophrenia; single nucleotide polymorphism; functional magnetic resonance imaging; deep canonically correlated sparse autoencoder

前言

精神分裂症是一种复杂的精神疾病,被认为是由环境和遗传因素的相互影响产生的。近年来,一些研究表明,单核苷酸多态性(Single Nucleotide Polymorphisms, SNP)等遗传因素在引起精神分裂症疾病中发挥重要作用,一些研究已开始探索精神分裂症研究的重要基因,已经发现许多与精神分裂症相关的潜在基因和基因组变异,例如染色体13q上的G72/G30基因位点^[1]、基因DISC1变异^[2]以及基因GRIK3、EFNA5、AKAP5的拷贝数变异^[3]。使用基因

组数据对精神分裂症患者进行分类已取得良好的效果^[4]。除了遗传学研究外,功能性磁共振成像(fMRI)一直是研究精神分裂症的一种有效工具,因为fMRI可以识别大脑内功能异常^[5],并且有成功应用的历史^[6-7]。近年来fMRI成像和遗传学的融合,以及fMRI和SNP数据的联合分析正成为一种更前沿的方法^[8]。然而,如何有效地组合fMRI和SNP数据是一个具有挑战性的问题。

典型相关分析(Canonical Correlation Analysis, CCA)是组合两种或更多种数据类型的通用方法^[9]。CCA通过寻找来自两种类型数据的变体的线性组合之间的最大相关性来找到彼此相关联的成分。在基因组和脑成像数据中,例如SNP和fMRI数据具有高维度但小样本的特点,因此CCA方法将面临过拟合的问题。为解决这个问题,CCA将稀疏惩罚项例如 L_1 范数^[10]或者 L_1 范数和 L_2 范数的组合(弹性网络)^[11]加入到传统的CCA模型来选择少量特征^[12]。此外,

【收稿日期】2019-10-18

【基金项目】长安大学研究生科研创新实践项目(300103002075)

【作者简介】李刚,副教授,工学博士,博士后,研究方向:模式识别、机器学习、多模态生物医学信息融合, E-mail: 15229296166@chd.edu.cn

CCA也增加了更有效的惩罚项,例如组稀疏CCA^[13]、结构化稀疏CCA^[14]、联合稀疏CCA^[15]和自适应稀疏CCA^[16]。然而,在线性CCA模型中没有考虑多种数据类型之间的复杂非线性关系,因此有人提出了深度CCA(DCCA)^[17]。相关自编码器用于优化两个数据集之间的典型相关性的组合。DCCA和相关自编码器专注于非线性信息转换,但忽略了有效的非线性降维。因此,如何同时实现有效的非线性信息转换和非线性降维是另一个挑战。

为了寻找两种类型数据之间复杂的非线性关系并解决小样本问题,本文提出深度典型相关稀疏自编码器(Deep Canonically Correlated Sparse Autoencoder, DCCSAE)方法。为了证明其应用,我们将DCCSAE应用于来自精神分裂症研究的SNP和fMRI数据集以获得特征表示,基于该特征表示将精神分裂症患者和健康人对照分类。实验结果表明,DCCSAE方法比现有的方法可以获得更准确的分类结果。

1 研究方法

1.1 深度典型相关分析

Andrew等^[17]设计了一种称为DCCA的典型相关分析的深度神经网络(Deep Neural Network, DNN),如图1所示。DCCA可以克服无法检测复杂非线性相关性的CCA限制。在DCCA中,两个DNN可以学习每个数据集的非线性表示。DCCA是通过最大化两个DNN输出的典型相关性获得的,如下所示:

$$(U^*, V^*) = \max_{W_f, W_g, U, V} \frac{1}{N} \text{tr}(U^T f(X) g(Y)^T V) \quad (1)$$

其中, $U^T \left(\frac{1}{N} f(X) f(X)^T + rI \right) U = I$, $V^T \left(\frac{1}{N} g(Y) g(Y)^T + sI \right) V = I$, $u_i^T f(X) g(Y)^T v_j = 0$, $i \neq j$ 。N表示数据总数,X和Y表示两个数据集的输入矩阵,I是单位矩阵,f(X)和g(Y)分别表示具有参数 W_f 和 W_g 的两个DNN的非线性表示,而U和V是最终输出的投影向量。

1.2 堆叠稀疏自编码器网络

自编码器(Autoencoder, AE)^[18]由编码层和解码层组成,它是一种前馈神经网络,与主成分分析一样,AE主要用于减少数据维度。

如图2所示,在编码层中,AE接收 $x \in R^p$ 作为输入并将其编码到隐藏层h以减小输入维度;在解码层中,降低维度的数据被解码为输出。输入矢量由下面的等式编码:

$$h = \sigma(Wx + b) \quad (2)$$

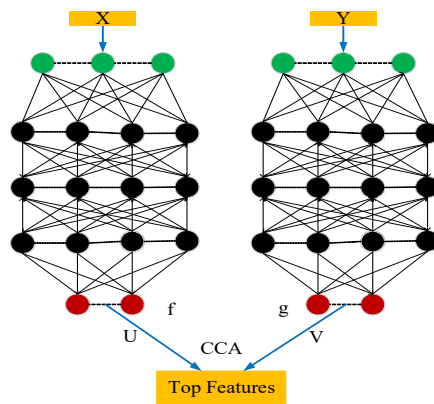


图1 深度典型相关分析图

Fig.1 Deep canonically correlation analysis

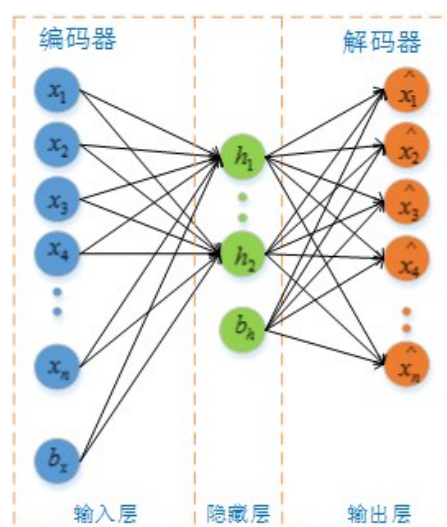


图2 单层自编码器图

Fig.2 Monolayer autoencoder

其中, σ 表示激活函数例如sigmoid、tanh, $W \in R^{n \times p}$ 是权重矩阵, $b \in R^n$ 是偏置向量。然后,使用解码方程对隐藏表达式进行解码以使数据 \hat{x} 接近原输入 x :

$$\hat{x} = \sigma(W'h + b') \quad (3)$$

其中, $W' \in R^{n \times n}$ 表示权重矩阵,而 $b' \in R^n$ 是偏差向量,输入 x 和输出 \hat{x} 之间的差异被称为重建误差。为了优化参数 W 、 W' 、 b 、 b' ,重建误差用作成本函数。对于单个训练示例,成本函数由下式表示:

$$J_{AE} = \frac{1}{2} \|\hat{x} - x\|^2 \quad (4)$$

对于多个训练示例(训练示例的数量是N),总成本函数由下式表示:

$$J_{AE} = \frac{1}{N} \sum_{i=1}^N \|\hat{x}^{(i)} - x^{(i)}\|^2 \quad (5)$$

过拟合是训练自编码器网络要面临的问题。因此,对成本函数添加权重惩罚项是解决过拟合的有效方法。带惩罚项的成本函数定义为:

$$J_{AE} = \frac{1}{N} \sum_{i=1}^N \frac{1}{2} \left\| \widehat{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)} \right\|^2 + \frac{\lambda}{2} (\|\mathbf{W}\|^2 + \|\mathbf{W}'\|^2) \quad (6)$$

当输入维度特别高或隐藏单元的数量很大时,在训练期间对隐藏单元施加稀疏性以发现输入的内部结构。如果神经元的输出是接近1的值,则认为神经元是激活的,而如果它的输出是接近0的值则被认为是非激活的。该隐藏单元的平均激活 j 被定义为:

$$\widehat{\rho}_j = \frac{1}{N} \sum_{i=1}^N h_j(\mathbf{x}^{(i)}) \quad (7)$$

为了实施稀疏性,我们限制 $\widehat{\rho}_j = \rho$, ρ 表示稀疏度目标,通常是接近0的很小的正数。我们添加相对熵作为惩罚因子以达到稀疏性,如下式所示:

$$J_{KL}(\rho \parallel \widehat{\rho}_j) = \sum_{j=1}^S \rho \log \frac{\rho}{\widehat{\rho}_j} + (1 - \rho) \log \frac{1 - \rho}{1 - \widehat{\rho}_j} \quad (8)$$

其中, S 表示隐藏层节点的数量。稀疏自编码器(Sparse Autoencoder, SAE)的总成本函数表示为:

$$J_{SAE} = \frac{1}{N} \sum_{i=1}^N \frac{1}{2} \left\| \widehat{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)} \right\|^2 + \frac{\lambda}{2} (\|\mathbf{W}\|^2 + \|\mathbf{W}'\|^2) + \beta J_{KL}(\rho \parallel \widehat{\rho}_j) \quad (9)$$

其中, β 表示稀疏性惩罚项。通过最小化成本函数,我们可以获得最优参数 \mathbf{W} 、 \mathbf{W}' 、 \mathbf{b} 和 \mathbf{b}' 。

自动编码器可以进一步堆叠在一起,以获取更多的信息。通过使用多个自编码器,提出了堆叠稀疏自编码器(Stack Sparse Autoencode, SSAE)^[19]。如图3所示,编码层连接到下一个SAE的输入层,以便更好地提取特征。

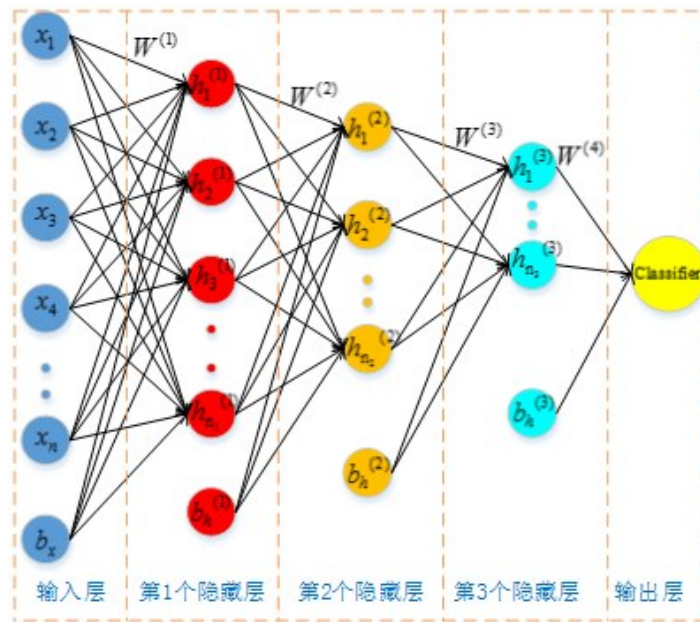


图3 堆叠稀疏自编码器图

Fig.3 Stacked sparse autoencoder

1.3 DCCSAE

在探索两类数据的非线性映射时,DCCA取得了良好的效果。SAE在寻求单个数据的非线性表示方面取得了巨大成功。然而DCCA无法实现有效的非线性降维,SAE无法探索跨模态数据之间的相关性。

我们将DCCA与SAE结合起来以获得两种数据类型的最佳表示,并提出一种新模型如图4所示,DCCSAE寻求两个数据集的深度网络表示,最大化两者之间的典型相关性,同时最小化稀疏自编码器的重建误差。DCCSAE定义为下式:

$$J_{DCCSAE} = \min_{\mathbf{W}_f, \mathbf{W}_g, \mathbf{U}, \mathbf{V}} - \frac{1}{N} \text{tr}(\mathbf{U}^T f(\mathbf{X}) g(\mathbf{Y})^T \mathbf{V}) +$$

$$\frac{\lambda}{N} \left\| \widehat{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)} \right\|^2 + \left\| \widehat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)} \right\|^2 + \alpha J_{KL}(\rho \parallel \widehat{\rho}_j) + \beta J_{KL}(\sigma \parallel \widehat{\sigma}_k) \quad (10)$$

其中, f 和 g 是用于为每个数据集提取非线性特征的DNN,同时对每个输入进行编码。 $\mathbf{U} = [u_1, \dots, u_L]$ 和 $\mathbf{V} = [v_1, \dots, v_L]$ 是将DNN输出投影到具有 L 个单位顶层的CCA方向。 $\widehat{\mathbf{x}}^{(i)}$ 和 $\widehat{\mathbf{y}}^{(i)}$ 分别表示输入 $\mathbf{x}^{(i)}$ 和输入 $\mathbf{y}^{(i)}$ 的重建。

$J_{KL}(\cdot)$ 定义为: $J_{KL}(\rho \parallel \widehat{\rho}_j) = \sum_{j=1}^S \rho \log \frac{\rho}{\widehat{\rho}_j} + (1 - \rho) \log \frac{1 - \rho}{1 - \widehat{\rho}_j}$ 。这里的 $\widehat{\rho}_j = \frac{1}{N} \sum_{i=1}^N h_j(\mathbf{x}^{(i)})$ 表示这个隐藏单位 j 的平均激活数,对于 $J_{KL}(\sigma \parallel \widehat{\sigma}_j)$ 也是类似的。

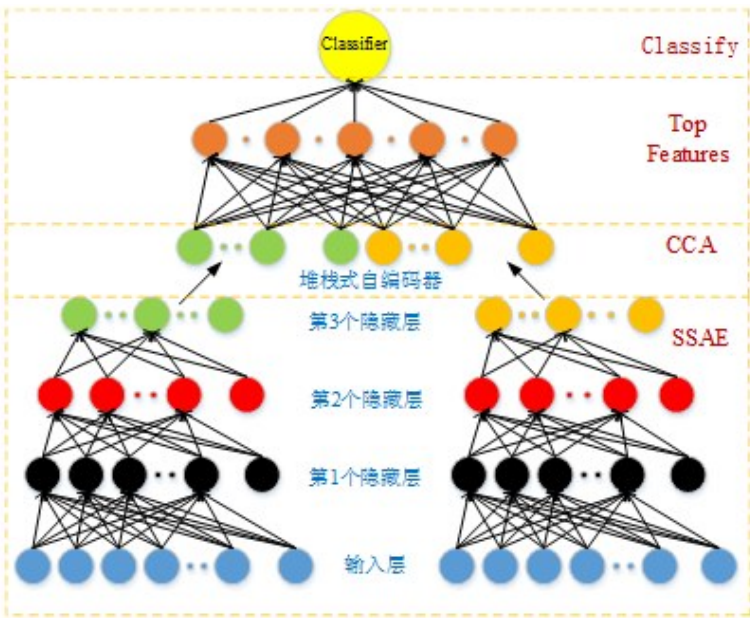


图4 深度典型相关稀疏自编码器图

Fig.4 Deep canonically correlated sparse autoencoder

2 实验结果

2.1 实验数据

本研究收集184名受试者(包括81名精神分裂症患者和103名健康对照者)的两个数据集(SNP和fMRI)。在传感器运动任务期间收集fMRI数据集,用SPM5预处理数据,并生成53×63×46体素的图像数据。我们进一步从116个感兴趣区域中提取41 236个体素,并基于AAL脑图谱进行分析。从每个受试者获得血液样品并

提取DNA,所有参与者的基因分型均在心智研究网络中进行,得到1140 419个SNP位点数据集。经质量控制,进一步筛选出722 177个SNP位点。

2.2 模型构建

两个DNN分别用于两个数据(SNP和fMRI)。每个DNN包含1个输入层和4个隐藏层。第一、二、三隐藏层是SSAE层,第四隐藏层是顶部特征层(DCCA层)。整个实验的工作流程如图5所示。

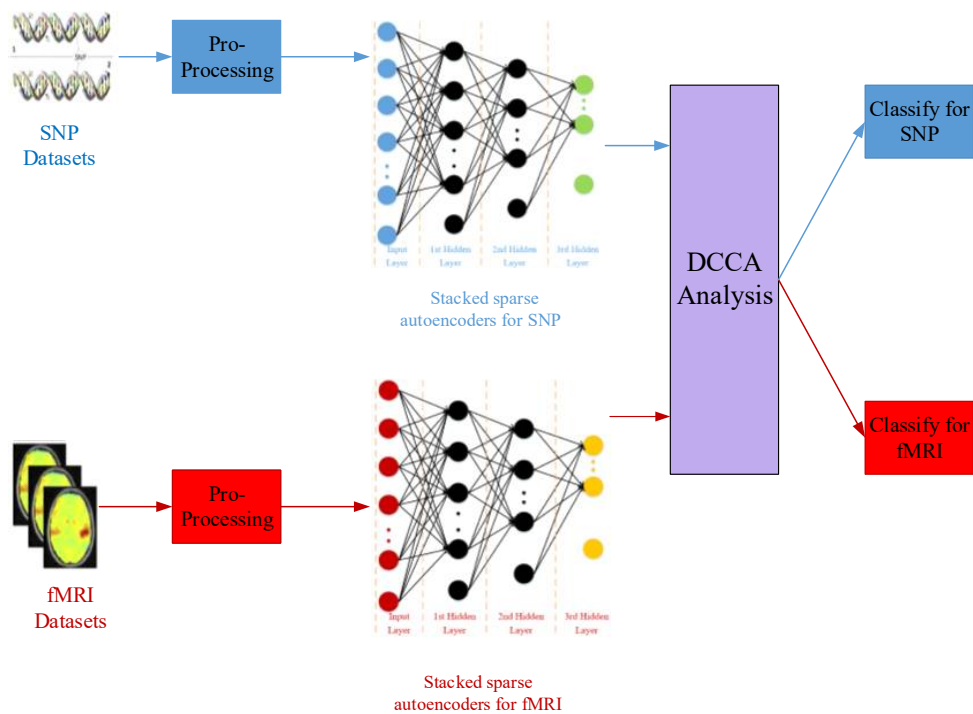


图5 模型结构图

Fig.5 Structure of experimental model

2.3 结果分析

为了获得最佳分类性能,我们在 DCCSAE 模型中添加了4个分类器,包括随机森林(Random Forest, RF)、神经网络(Neural Network, NN)、逻辑回归(Logist)和支持向量机(Support Vector Machine, SVM)。如表 1 和表 2 所示,采用 DCCSAE+SVM 时,分类性能最佳。

表 1 在 SNP 数据集上的分类性能%
Tab.1 Classification performance on SNP dataset %

模型	准确率	精确度	召回率	F1 值
DCCSAE+RF	92.79±0.52	92.21±0.49	91.84±0.43	91.56±0.65
DCCSAE+NN	94.47±0.35	93.88±0.29	93.63±0.33	92.92±0.41
DCCSAE+Logist	89.66±0.58	88.75±0.52	88.68±0.54	89.56±0.61
DCCSAE+SVM	95.65±0.14	93.74±0.18	94.57±0.23	93.82±0.30

表 2 在 fMRI 数据集上的分类性能%
Tab.2 Classification performance on fMRI dataset %

模型	准确率	精确度	召回率	F1 值
DCCSAE+RF	74.94±0.65	74.32±0.61	73.68±0.67	73.53±0.59
DCCSAE+NN	76.86±0.53	75.97±0.58	76.32±0.55	75.85±0.63
DCCSAE+Logist	78.23±0.39	77.57±0.43	78.03±0.47	76.95±0.45
DCCSAE+SVM	80.53±0.23	78.69±0.29	79.76±0.32	78.85±0.41

在前面的实验中,我们已经证明 DCCSAE+SVM 具有最佳的分类性能,因此在多个模型的比较中,将 SVM 分类器用于每个模型。如图 6 所示,SNP 和 fMRI 数据集的分类结果清楚地表明 DCCSAE 模型优于其他分类器。DCCA(深度 CCA)模型比 CCA、SCCA(稀疏 CCA)和 KCCA(核 CCA)获得更高的准确度,而所提出的 DCCSAE 模型实现了最佳分类,SNP 数据集的准确率为

95.65%,fMRI 数据集的准确率为 80.53%。根据精确度(正确识别精神分裂症患者的得分)和召回率(预测精神分裂症患者对所有样本的得分),DCCSAE 在精确度和召回率之间也得到了良好的权衡。此外,从最高的 F1 分数(精确度和召回率的加权调和平均值),我们可以得出 DCCSAE 分类模型在精神分裂症分类中优于其他模型。

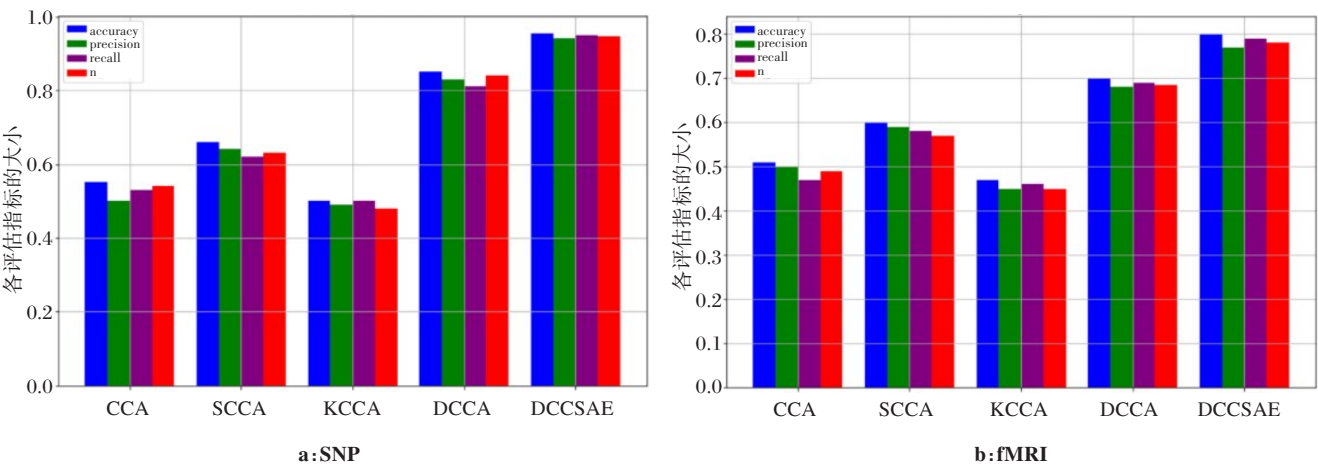


图 6 不同模型在两个数据集上的分类表现
Fig.6 Classification performances of different models on two datasets

SNP 数据的 ROC 曲线如图 7a 所示,fMRI 数据的 ROC 曲线如图 7b 所示,我们发现所提出的 DCCSAE 模型在 SNP 数据集上的 AUC 得分是 0.984,在 fMRI 数据集上的 AUC 得分是 0.953,对比其他的模型,这

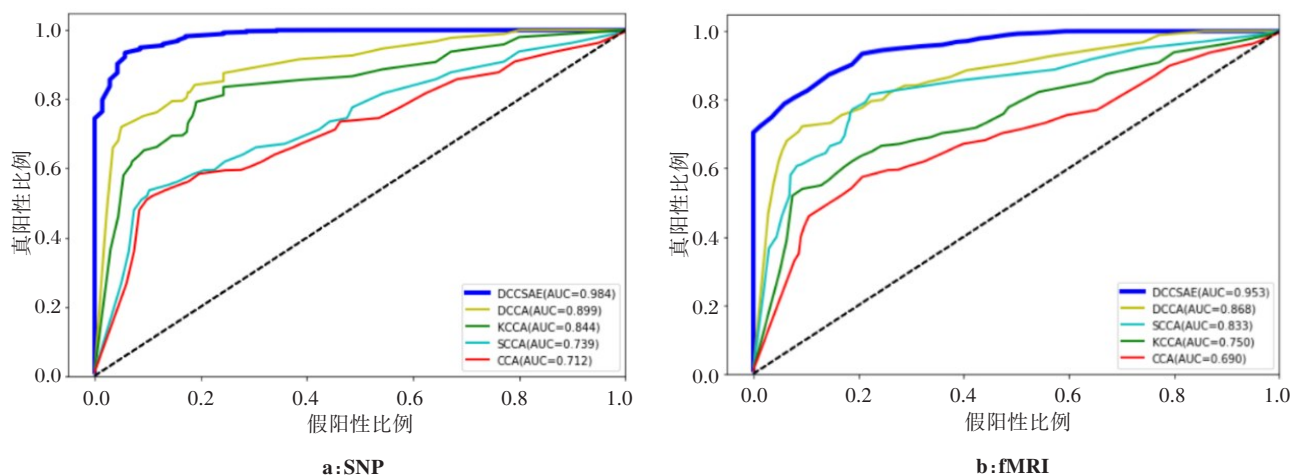


图7 不同模型在两个数据集上的ROC曲线

Fig.7 Receiver operating characteristic curves of different models on different datasets

是最高的AUC得分。这些表明DCCSAE模型在精神分裂症的分类和预测上达到了很好的效果。

3 结语

本文提出用于精神分裂症分类的DCCSAE模型,并将所提出的DCCSAE模型应用于SNP数据和fMRI数据以测试其性能。为了获得每层的初始化,逐层训练堆叠的自动编码器,并在隐藏层中添加稀疏性。然后DCCA用于连接两种类型的数据,我们通过反向传播来微调网络。最后,我们使用所提出的模型以及包括CCA、SCCA、KCCA和DCCA在内的多种方法来分析SNP数据和fMRI数据以进行性能评估。就若干指标而言,与其他4种CCA相关模型相比,所提出的DCCSAE模型在精神分裂症分类中具有更好的性能。因此,该模型在疾病分类和预测方面显示出巨大的前景。

【参考文献】

- [1] BADNER J, GERSHON E. Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia[J]. Mol Psychiatr, 2002, 7(4): 405-411.
- [2] CALLICOTT J H, STRAUB R E, PEZAWAS L, et al. Variation in DISC1 affects hippocampal structure and function and increases risk for schizophrenia[J]. Proc Natl Acad Sci U S A, 2005, 102(24): 8627-8632.
- [3] WILSON G M, STEPHANE F, VIKRAMJIT C, et al. DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling[J]. Hum Mol Genet, 2006, 15(5): 743-749.
- [4] VIVIAN-GRIFFITHS T, BAKER E, SCHMIDT K M, et al. Predictive modeling of schizophrenia from genomic data: comparison of polygenic risk score with kernel support vector machines approach[J]. Am J Med Genet B, 2019, 180(1): 80-85.
- [5] MEDA S A, Bhattarai M, Morris N A, et al. An fMRI study of working memory in first-degree unaffected relatives of schizophrenia patients

[J]. Schizophr Res, 2008, 104(1-3): 85-95.

- [6] RASHID B, ARBABSHIRANI M R, DAMARAJU E, et al. Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity[J]. NeuroImage, 2016, 134(4): 645-657.
- [7] YANG H, HE H, ZHONG J. Multimodal MRI characterisation of schizophrenia: a discriminative analysis[J]. Lancet, 2016, 388(1): 36.
- [8] LIU J, PEARLSON G, RUANO G, et al. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA[J]. Hum Brain Mapp, 2010, 30(1): 241-255.
- [9] HOTELLING H. Relations between two sets of variates [J]. Biometrika, 1936, 28(3-4): 321-377.
- [10] CAO K A, MARTIN P G, ROBERT-GRANIÉ C, et al. Sparse canonical methods for biological data integration: application to a cross-platform study[J]. BMC Bioinformatics, 2009, 10(1): 1-17.
- [11] WITTEN D M, TIBSHIRANI R J. Extensions of sparse canonical correlation analysis with applications to genomic data[J]. Stat Appl Genet Mol, 2009, 8(1): 1-27.
- [12] 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.
- [13] LIN D, CALHOUN V D, WANG Y P. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis [J]. Med Image Anal, 2014, 18(6): 891-902.
- [14] DU L, HUANG H, YAN J, et al. Structured sparse canonical correlation analysis for brain imaging genetics: an improved GraphNet method[J]. Bioinformatics, 2016, 32(10): 1544-1551.
- [15] 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.
- [16] 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 Trans Biomed Eng, 2018, 65(2): 390-399.
- [17] ANDREW G, ARORA R, BILMES J, et al. Deep canonical correlation analysis[C]. International Conference on International Conference on Machine Learning, 2013: 1047-1255.
- [18] BALDI P. Autoencoders, unsupervised learning and deep architectures [C]. International Conference on Unsupervised & Transfer Learning Workshop, 2011: 37-50.
- [19] XU J, XIANG L, LIU Q, et al. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images[C]. IEEE International Symposium on Biomedical Imaging, 2014: 999-1002.

(编辑:陈丽霞)