

基于统计特性随机森林算法的特征选择

宋 源, 梁雪春*, 张 然

(南京工业大学 自动化与电气工程学院, 南京 211816)

(* 通信作者电子邮箱 hjsdym@163.com)

摘 要: 针对由静息态功能磁共振成像(R-fMRI)得到的脑功能连接矩阵数据运用传统特征选择方法处理的结果, 存在特征冗余, 无法确定最终特征维数等问题, 提出一种全新的特征选择算法。该算法在随机森林(RF)算法中结合统计特性, 根据袋外数据的分类效果得到保留的特征, 并将其运用在对精神分裂患者与正常被试者的识别实验中。实验结果表明, 与传统的主成分分析(PCA)方法相比, 该算法可以有效保留重要特征, 提高识别精度, 且保留的特征具有很好的医学解释性。

关键词: 随机森林; 统计特性; 静息态功能磁共振成像; 脑功能连接矩阵

中图分类号: TP391.4 **文献标志码:** A

Feature selection based on statistical random forest algorithm

SONG Yuan, LIANG Xuechun*, ZHANG Ran

(College of Automation and Electrical Engineering, Nanjing Technology University, Nanjing Jiangsu 211816, China)

Abstract: Focused on the traditional methods of feature selection for brain functional connectivity matrix derived from Resting-state functional Magnetic Resonance Imaging (R-fMRI) have feature redundancy, cannot determine the final feature dimension and other problems, a new feature selection algorithm was proposed. The algorithm combined Random Forest (RF) algorithm in statistical method, and applied it in the identification experiment of schizophrenic and normal patients, according to the features are obtained by the classification results of out of bag data. The experimental results show that compared to the traditional Principal Component Analysis (PCA), the proposed algorithm can effectively retain important features to improve recognition accuracy, which have good medical explanation.

Key words: Random Forest (RF); statistical property; Resting-state functional Magnetic Resonance Imaging (R-fMRI); brain functional connectivity matrix

0 引言

功能磁共振成像(functional Magnetic Resonance Imaging, fMRI)技术近年来得到研究者的广泛关注^[1]。该技术通过探测大脑中的血氧饱和和变化来达到无侵入性观察大脑活动变化的目的。其中静息态功能磁共振成像(Resting-state functional Magnetic Resonance Imaging, R-fMRI)反映的是占大脑活动90%能量的自发性活动, 这些自发性活动被认为与精神分裂症、抑郁症、癫痫、阿尔茨海默病等一些大脑疾病相关^[2]。在针对R-fMRI的研究中, 研究者一般关注反映不同脑区间神经生理活动相关的功能连接特性。这些特性往往能反映正常被试者与病人在脑功能连接上的区别, 但是由于fMRI数据量庞大, 提取的特征往往存在维数过大、数据重叠等问题^[3]。

为解决上述问题, 国内外学者分别采用多种方法对特征指标进行优化选取, 其中最常用的是t检验, 其根据变量的p值将特征排序。然而这种单变量特征排序没有考虑变量间可能存在的相互作用。因此Breiman等提出根据随机森林(Random Forest, RF)算法去建立特征重要度表^[4], 这种方法虽然能给出较准确的特征排序, 但却无法解释结果, 更为重要

的是无法给出将特征确定为重要特征的阈值。因此本文在该算法基础上提出一种综合统计特性的特征排序方法。该算法能综合假设检验及机器学习算法两者的优点, 并准确给出需要保留的特征。

本文首先使用改进的随机森林算法设置 $p < 0.05$ 为阈值, 删除一些在统计结果中“不重要”的特征, 随后将保留的特征作为支持向量机(Support Vector Machine, SVM)的输入特征, 并利用测试集进行识别, 通过最终识别结果来验证效果。实验仿真结果证明, 该算法能有效减少特征数目, 且选出的指标具有很好的分类效果。

1 功能性磁共振成像数据特征提取

本文主要研究的是精神分裂患者(Schizophrenia, Sz)与正常被试者之间的特征差异, 而精神分裂作为一种脑部无明显结构异常的疾病, 其病因原因还未知。随着研究者的深入研究发现, 一些精神类疾病可能与默认脑区的活动异常有关^[5]。

本文选用解剖学模板(Anatomical Automatic Labeling, AAL)将全脑分为90个脑区^[6], 并提取每个脑区的平均体素

收稿日期: 2014-11-21; 修回日期: 2015-01-07。

基金项目: 国家自然科学基金资助项目(51205185); 江苏省普通高校研究生科研创新计划项目(KYLX_0754)。

作者简介: 宋源(1989-), 男, 江苏南京人, 硕士研究生, 主要研究方向: 复杂系统预测及建模; 梁雪春(1969-), 女, 江苏南京人, 教授, 博士, 主要研究方向: 复杂系统预测及建模; 张然(1989-), 女, 安徽淮南人, 硕士研究生, 主要研究方向: 复杂系统预测及建模。

时间序列,时间序列矩阵如下:

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,N} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,N} \\ \vdots & \vdots & & \vdots \\ d_{90,1} & d_{90,2} & \cdots & d_{90,N} \end{bmatrix} \quad (1)$$

其中: $d_{i,j}$ 为在 j 时刻第 i 脑区的所有体素的平均值, N 为时间序列长度。根据医学先验知识,选择与精神疾病相关脑区的 20 个脑区作为感兴趣区^[7]。通过计算 20 个脑区平均体素时间序列皮尔森相关系数去构建功能连接网络,公式如下:

$$r_{ij} = \frac{\sum_{t=1}^T [d_i(t) - \bar{d}_i] \cdot [d_j(t) - \bar{d}_j]}{\sqrt{\sum_{t=1}^T [d_i(t) - \bar{d}_i]^2} \sqrt{\sum_{t=1}^T [d_j(t) - \bar{d}_j]^2}} \quad (2)$$

其中: d_i 和 d_j 分别表示体素 i 和体素 j 的时间序列, \bar{d}_i 和 \bar{d}_j 分别表示时间序列 d_i, d_j 的均值,这样就得到脑区之间的功能连接系数矩阵:

$$R = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,20} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,20} \\ \vdots & \vdots & & \vdots \\ r_{20,1} & r_{20,2} & \cdots & r_{20,20} \end{bmatrix} \quad (3)$$

这是一个 20×20 的对称矩阵,选取上三角元素构成一个 1×190 的特征向量。对于只有小样本数据的 Sz 患者数据,这样的特征需要经过降维使用。

2 基于统计特性的 RF 的特征重要度排序

2.1 随机森林特征重要度

随机森林(RF)模型是由决策回归树发展而成^[8],随机森林常常会生成几百棵树,每棵树的数据都是由自助抽样法(Bootstrap Sampling)从定义为集合 B 的袋中抽出,余下的不出现在训练样本中的袋外数据(Out-Of-Bag, OOB)样本定义为集合 \bar{B} ,定义 C 为 B 的集合,而 \bar{C} 为 \bar{B} 的集合。假设 $X^{n \times p}$ 矩阵为 n 个具有 p 个特征的被试数据集, y 为一个 n 维标签向量,每一个值代表所对应的被试所属类别。随机森林算法通过重新排列特征前后的分类误差计算特征的重要度,该算法中每一个特征 x_j 都对应着一组取值重新排列的特征置换测试。通过在 OOB 测试集中比较使用原始特征及置换后的随机重排特征的分类误差率来衡量特征的重要度。当重要的特征被随机重排特征置换时,其区分度会下降,即 OOB 分类误差率上升。当建立 T 棵树时就有 T 个 OOB 集作为测试集。因此有定义特征重要度指标 J_a , 公式如下:

$$J_a(x_j) = \frac{1}{T} \sum_{B_k \in \bar{C}} \frac{1}{|B_k|} \left(\sum_{i \in B_k} I(h_k^j(i) \neq y_i) - I(h_k(i) \neq y_i) \right) \quad (4)$$

其中: y_i 为在第 i 个 OOB 中的分类标签, I 为示性函数, $h_k(i)$ 为通过数据集 B_k 预测的样本 i 的分类标签, $h_k^j(i)$ 为置换特征 x_j 后的分类标签。

2.2 基于统计特性的随机森林重要度排序

虽然 J_a 可以综合变量间的关系体现出单个变量的重要度,但是其结果缺少解释性,特别是该算法并不能定义一个确切的阈值去突出其统计学意义,因此本文将在 Breiman 提出

的 J_a 指标基础上结合统计测试来评价特征的重要度。

针对文本研究的识别精神分裂与正常被试的二分类问题,对于每一个在 OOB 中的数据,其预测结果可分为以下四类:正确预测类 1 即正确识别 Sz 患者(TP),正确预测类 0 即正确识别正常被试者类(TN),错误预测类 1 即正常被试者被识别为 Sz 患者(FP)以及错误识别类 0 即 Sz 患者被识别为正常被试者(FN),统计这些值可以估计出整个 OOB 上的预测结果,同样该方法对于估计特征 x_j 对预测结果的影响也适用。其特征影响可以用一个 4×2 的列联表表出,公式如下:

	x_j	\tilde{x}_j
TN	$s(0,0)$	$\tilde{s}^j(0,0)$
FP	$s(0,1)$	$\tilde{s}^j(0,1)$
FN	$s(1,0)$	$\tilde{s}^j(1,0)$
TP	$s(1,1)$	$\tilde{s}^j(1,1)$

(5)

其中

$$s(l_1, l_2) = \sum_{\bar{B}_k \in \bar{B}_i \in \bar{B}_k} I(y_i = l_1 \text{ and } h_k(i) = l_2) \quad (6)$$

而 $\tilde{s}^j(l_1, l_2)$ 的定义与上文的 h_k^j 的定义类似。

本文使用皮尔逊 χ^2 检验评估原始的 x_j 与替换过的 \tilde{x}_j 频数是否具有显著差异,若在 p 值较小时,拒绝 0 假设。那该 $p_{\chi^2}(x_j)$ 意味着交换变量 x_j 对投票结果具有重要影响。因为几个特征的重要性都是通过同一组数据进行评估, p 值必须正确反映多次测试结果。本文使用 Benjamini-Hochberg 系数^[9]去控制错误发生率。设 $p_{\chi^2}^{thr}(x_j)$ 为 $p_{\chi^2}(x_j)$ 经过错误修正的值。定义如下:

$$J_{\chi^2}(x_j) = p_{\chi^2}^{thr}(x_j) \quad (7)$$

基于统计特征的 J_{χ^2} 值,虽然与 J_a 值类似,但有如下区别:1) J_a 通过一个指标衡量所有错误,这可能会无法反映一些非平衡数据的特征;2) J_a 在 OOB 样本集中估计分类错误,而 J_{χ^2} 是通过改变样本集去估计分布。最为重要的是 J_{χ^2} 值更容易解释特征的重要度,较低的 J_{χ^2} 值意味着该变量对最终分类的结果影响越不显著,因此可以通过设定 $p < 0.05$,来设定阈值去拒绝不重要的指标^[10]。

3 实验结果与指标分析

3.1 实验数据

本文研究所使用的数据来源于美国哈特福德奥林神经精神病学研究中心,被试为 36 名健康者和 26 名精神分裂症患者,共 62 名被试者参加了静息态功能性磁共振成像实验,被试者均知情同意。磁共振仪器为德国西门子公司的 3T 磁共振成像系统,采用平面回波成像序列采集静息态 fMRI 数据。扫描参数为:重复时间(Repetition Time, TR)为 1500 ms,回波时间(Echo Time, TE)为 27 ms,视野(Field Of Vision, FOV)为 $24 \text{ cm} \times 24 \text{ cm}$,翻转角(Flip Angle, FA)为 90° ,矩阵为 53×63 ,层厚 4 mm,层距 1 mm,共 46 层。使用 SPM8 完成时间层校正、头动校正、滤波、平滑等操作,并去除头动、脑脊液、全脑信号等协变量。

3.2 实验结果

J_{χ^2} 是选择 RF 重要特征指标中关键变量的判断标准,目的是减少用于最终分类时的特征指标。本文在将通过被试数据测试这一标准与相关变量选择的匹配度。

本文算法根据所有特征的重要性对特征排序,根据统计特性,当一个特征的 p 值低于0.05时,则该指标被认为是显著重要的。图1表明了特征在RF不同规模生成树以及 m 值中的重要性,这里的 m 值对应的是RF在生成树时随机选择特征的候选数。根据文献[11],在高维度的数据集中,RF在生成500棵树时可获得良好的效果,在生成10000棵树时能够得到稳定的特征选择。因此,本文用RF生成500棵树、生成10000棵树以及 m 值分别为10和110时建模,结果如图1所示。

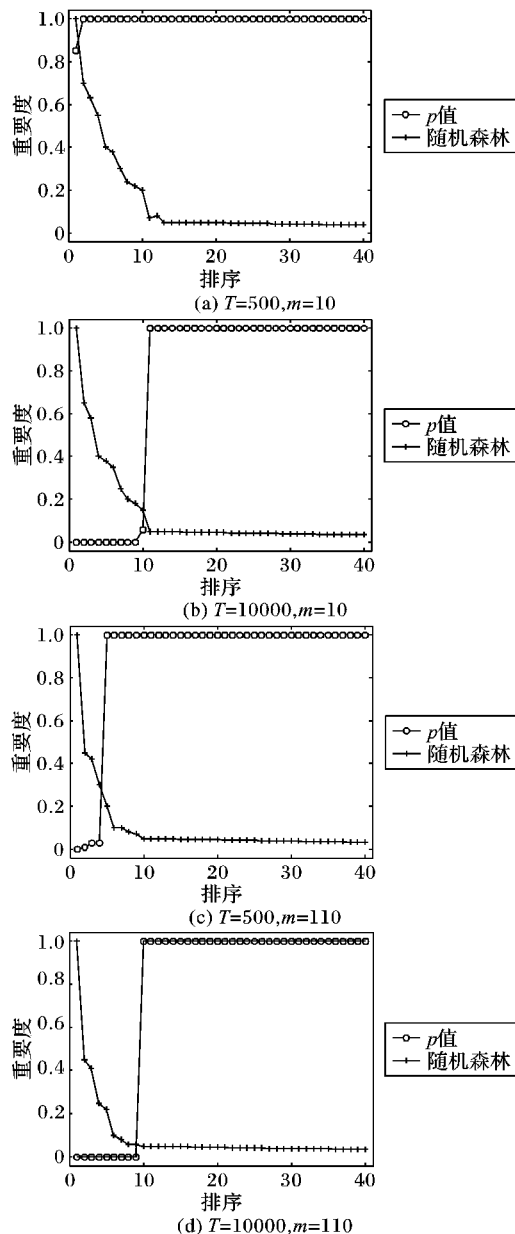


图1 重要度和 p 值随参数变化

图1的结果表明,特征指标的重要度与其统计特性具有一致性,重要度高的特征指标 p 值则较低,表明该特征指标在统计结果中也显著重要。针对排名前列的10个信息量大的特征出现在 J_a 和 J_{x2} 的排名前列,原始的Breiman的 J_a 无法给出一个明确的标准判别该特征是否相关,而本文提出的 J_{x2} 排序可以更好地判别相关特征与不相关特征。不足的是, J_{x2} 的排序需要生成大量的树获得稳定的结果,如图1(a)所示,

当生成少量的树($T = 500$)时, J_{x2} 无法判断特征是否显著重要,不过特征排序仍然是正确的排名。如图1(c)所示,当生成树较少时,增加 m 值同样可以增加辨别特征显著相关的能力,但是效果不如生成更多的树明显。实验结果表明图1(d)所示的效果最好,这也证明RF生成树的规模越大,特征选择的效果越好。

因此在下文的分析中,本文选用 $T = 10000, m = 110$ 这组参数。

3.3 实验结果

本文选择“留一法”分别研究不同算法选择出的特征对识别精度的影响。为量化识别效果,本文选用识别率 GR 、灵敏度 SS 、特异度 SC 如下:

$$GR = \frac{TP + TN}{TP + FN + TN + FP} \quad (8)$$

$$SS = \frac{TP}{TP + FN} \quad (9)$$

$$SC = \frac{TN}{TN + FP} \quad (10)$$

其中: TP 、 TN 、 FP 和 FN 分别代表正确识别Sz患者的人数、正确识别正常被试的人数、正常被试被识别为Sz患者的人数和Sz患者被识别为正常被试的人数。这些指标能反映每类的分类精度,特别是针对非平衡数据时,这类指标比一般的分类精度具有更好的评价性能。

为了克服单一分类器带来的识别误差,在最终的识别中,本文使用SVM作为最终的分类器。统计特性改进的随机森林算法(Statistical Random Forest, SRF)从原始190个特征选出10个特征作为最终SVM输入,与传统的主成分分析法(Principal Component Analysis, PCA)(选择95%的能量成分)及 t 检验($p < 0.05$)选择的特征对比的识别结果如表1所示。

表1 不同特征选择的分类结果 %

算法	识别率	灵敏度	特异度
PCA + SVM	75.8	76.9	75.0
t + SVM	74.2	76.9	72.2
SRF + SVM	85.5	84.6	86.1

实验数据表明与传统的PCA算法及 t 检验方法比较,本文算法在识别率、灵敏度和特异度上都有显著提高。

4 结语

本文提出一种统计方法与机器学习算法随机森林相结合的方法,并将该方法运用在精神分裂疾病的识别运用中,该方法能有效提取出具有显著差异的特征,最终保留10个特征,将该特征放回到原始功能连接矩阵中,就可得到最终保留的矩阵功能连接特征,这些功能连接涉及右侧外侧顶叶、尾状核左侧额上回、左侧海马旁回、右侧颞下回、后扣带回/楔前叶。这些脑区在目前的研究中被划分为默认脑区,临床医学研究表明默认脑区与自发的精神活动高度相关^[12]。这和本文的研究相符合,然而精神分裂的种类众多,不同症状的患者其连接特征也不尽相同。在以后的研究中,可以将本文中针对二分类问题的算法扩展到多分类问题。

(下转第1466页)

- segmentation methods based on graph cuts[J]. *Acta Automatica Sinica*, 2012, 38(6): 911–922. (刘松涛, 殷福亮. 基于图割的图像分割方法及其新进展[J]. *自动化学报*, 2012, 38(6): 911–922.)
- [2] JIANG S, YI F, TANG L, *et al.* Tumor extraction method of MRI cerebral image based on graph cuts[J]. *Computer Engineering*, 2010, 36(7): 217–219. (蒋世忠, 易法令, 汤浪平, 等. 基于图割的MRI脑部图像肿瘤提取方法[J]. *计算机工程*, 2010, 36(7): 217–219.)
- [3] SHANG Y, WANG N, WANG H. Medical object extraction model based on regional energy minimization and active contour model[J]. *Application Research of Computers*, 2012, 29(7): 2715–2718. (尚岩峰, 汪宁, 汪辉. 基于区域能量最小和主动轮廓模型的医学目标提取[J]. *计算机应用研究*, 2012, 29(7): 2715–2718.)
- [4] BOYKOV Y, JOLLY M. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images[C]// *Proceedings of the 8th IEEE International Conference on Computer Vision*. Piscataway: IEEE, 2001: 105–112.
- [5] BAUER C, POCK T, SORANTIN E, *et al.* Segmentation of interwoven 3D tubular tree structures utilizing shape priors and graph cuts[J]. *Medical Image Analysis*, 2010, 14(2): 172–184.
- [6] LYU Q, GAO J, GAO X. Strong scattering objects segmentation based on graph cut and mean shift algorithm from SAR images[J]. *Journal of Computer Applications*, 2014, 34(7): 2018–2022. (吕倩, 高君, 高鑫. 基于图割及均值漂移的合成孔径雷达图像强散射目标分割[J]. *计算机应用*, 2014, 34(7): 2018–2022.)
- [7] SHANG Y, WANG H, WANG N, *et al.* Three dimensional vessel extraction model based on tubular characters and active contour model[J]. *Journal of Image and Graphics*, 2013, 18(3): 290–298. (尚岩峰, 汪辉, 汪宁, 等. 管状特性和主动轮廓的3维血管自动提取[J]. *中国图象图形学报*, 2013, 18(3): 290–298.)
- [8] XU X, DING S, SHI Z, *et al.* New theories and methods of image segmentation[J]. *Acta Electronica Sinica*, 2010, 28(2): 76–82. (许新征, 丁世飞, 史忠植, 等. 图像分割的新理论和新方法[J]. *电子学报*, 2010, 38(2): 76–82.)
- [9] PENG B, ZHANG L, ZHANG D, *et al.* Image segmentation by iterated region merging with localized graph cuts[J]. *Pattern Recognition*, 2011, 44(10): 2527–2538.
- [10] ZHAO H. Research on and implementation of the algorithm of CT-based cardiac coronary artery segmentation[D]. *Shenyang: Northeastern University*, 2008. (赵宏伟. 基于CT数据的冠脉提取和细化方法的研究和实现[D]. 沈阳: 东北大学, 2008.)
- [11] GROSCEGEORGE D, PETITJEAN C, DACHER J N, *et al.* Graph cut segmentation with a statistical shape model in cardiac MRI[J]. *Computer Vision and Image Understanding*, 2013, 117(9): 1027–1035.
- [12] LERME N, LETOCART L, MALGOUYRES F. Reduced graphs for min-cut/max-flow approaches in image segmentation[J]. *Electronic Notes in Discrete Mathematics*, 2011, 37: 63–68.
- [13] ZHANG S, DONG J, SHE L. The methodology of evaluating segmentation algorithms on medical image[J]. *Journal of Image and Graphics*, 2009, 14(9): 1872–1880. (张石, 董建威, 余黎煌. 医学图像分割算法的评价方法[J]. *中国图象图形学报*, 2009, 14(9): 1872–1880.)

(上接第1461页)

参考文献:

- [1] WISNER K M, ATLURI G, LIM K O, *et al.* Neurometrics of intrinsic connectivity networks at rest using fMRI: retest reliability and cross-validation using a meta-level method[J]. *NeuroImage*, 2013, 32(6): 236–251.
- [2] MEDA S A, GILL A, STEVENS M C, *et al.* Differences in resting-state functional magnetic resonance imaging functional network connectivity between schizophrenia and psychotic bipolar probands and their unaffected first-degree relatives[J]. *Biological Psychiatry*, 2012, 71(10): 881–889.
- [3] ZHI L, LI Y, ZHAO S. Feature extraction of fMRI data based on discrete wavelet transform[J]. *Chinese Journal of Medical Imaging Technology*, 2010, 26(6): 1151–1154. (支联合, 李玉晓, 赵书俊. 基于离散小波变换的fMRI数据特征提取[J]. *中国医学影像技术*, 2010, 26(6): 1151–1154.)
- [4] LYU B, WANG H. High-dimensional data visualization based on random forest[J]. *Journal of Computer Applications*, 2014, 34(6): 1613–1617. (吕兵, 王华珍. 基于随机森林的高维数据可视化[J]. *计算机应用*, 2014, 34(6): 1613–1617.)
- [5] FAN M, WU F, ZHENG H. Study of cognitive function in ultra-high risk subjects and first episode patients with schizophrenia[J]. *Journal of Clinical Psychiatry*, 2014, 24(1): 11–13. (范敏珍, 吴逢春, 郑洪波. 精神分裂症超高危人群及首次发病患者的认知功能研究[J]. *临床精神医学杂志*, 2014, 24(1): 11–13.)
- [6] 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.
- [7] SURGULADZE S A, CHU E M, MARSHALL N, *et al.* Emotion processing in schizophrenia: fMRI study of patients treated with risperidone long-acting injections or conventional depot medication[J]. *Journal of Psychopharmacology*, 2011, 25(6): 722–733.
- [8] LIN C, PENG G. Application of random forest on selecting evaluation index system for enterprise credit assessment[J]. *Journal of Xiamen University: Natural Science*, 2007, 46(2): 199–203. (林成德, 彭国兰. 随机森林在企业信用评估指标体系确定中的应用[J]. *厦门大学学报: 自然科学版*, 2007, 46(2): 199–203.)
- [9] BENJAMINI Y, HOCHBERG Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing[J]. *Journal of the Royal Statistical Society, Series B: Methodological*, 1995, 57: 289–300.
- [10] STIGLER S. Fisher and the 5% level[J]. *Chance*, 2008, 21(4): 12.
- [11] KALOUSIS A, PRADOS J, HILARIO M. Stability of feature selection algorithms: a study on high-dimensional spaces[J]. *Knowledge and Information Systems*, 2007, 12(1): 95–116.
- [12] LI Y, WANG E, ZHANG H, *et al.* Default mode network in primary insomnia: measured with resting state functional MRI[J]. *Chinese Journal of Medical Imaging*, 2014, 22(7): 481–486. (李永丽, 王恩锋, 张红菊, 等. 原发性失眠患者默认网络神经功能的静息态MRI研究[J]. *中国医学影像学杂志*, 2014, 22(7): 481–486.)