

基于随机森林算法的特征选择及在 fMRI 数据中的应用

袁晓龙, 梅 雪, 黄嘉爽, 杨 骥

(南京工业大学 自动化与电气工程学院, 江苏 南京 211816)

摘 要: fMRI 数据是典型的高维小样本数据, 如何从高维数据中提取和选择重要的特征是正确分类识别的关键. 结合 fMRI 数据自身特点, 提出了一种应用随机森林算法进行特征选择的方法, 以随机森林分类精度为准则函数对特征进行重要性度量的方式实现特征选择. 将本方法应用于健康者和精神分裂症患者的识别中, 通过计算每个特征对分类的贡献度, 优选出贡献度大的特征用于分类识别, 同时根据重要特征的序号定位到相应脑区, 给临床诊断提供客观参照. 实验结果表明, 该方法具有较好的效果.

关键词: 分类识别; 特征提取与选择; 随机森林; fMRI 数据

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2014)08-0132-04

Feature Selection Based on Random Forest Algorithm and the Applications in fMRI Data

YUAN Xiao-long, MEI Xue, HUANG Jia-shuang, YANG Ji

(School of Automation & Electrical Engineering, Nanjing University of Technology, Nanjing 211816, China)

Abstract: fMRI data is typical small sample data with high-dimension, how to extract and select important features from high-dimensional data is the key for classification and recognition. Combined with the characteristics of fMRI data, a method of feature selection using Random Forest algorithm is proposed in this paper. To achieve the feature selection, Random Forest classification accuracy as a criterion to estimate the importance of features. The method was applied to healthy controls and patients with schizophrenia recognition, by calculating the contribution of each feature for classification to choose the features with greater contribution for classification and recognition, and locate to the corresponding brain regions according the order numbers of important features, providing an objective reference for clinical diagnosis. Experimental results show this method has a better performance.

Key words: classification and recognition; feature extraction and selection; random forest; fMRI data

1 引言

人脑是一个复杂的进化产物, 是控制运动、产生感觉及实现高级脑功能的高级神经中枢. 随着医学影像技术和神经科学的飞速发展, 人类可以借助于先进的仪器对大脑进行无损的成像研究. 功能性磁共振成像(functional magnetic resonance imaging, fMRI)技术则是一种广泛采用的脑成像方法. 然而, 采集获得的 fMRI 数据具有小样本超高维的特点^[1], 面对如此小的样本, 从超高维的数据中选择和

提取特征是提取有价值信息的前提, 也是正确分类的关键. 近年来, 很多研究者对 fMRI 数据的降维和特征选择提取进行了研究. 主要有通过主成分分析(Principle Component Analysis, PCA)、线性判别分析(Linear Discriminant Analysis, LDA)和投影寻踪(Projection Pursuit, PP)等等^[2-5]. 通过研究上述数据降维和特征选择方法, 发现上述任一种方法只能选择任意个重要的成份作为特征向量, 不能根据数据集的自身特点选出适合分类的特征, 而且在训练样本不足的情况下, 不能够对高维数据很好地

收稿日期: 2013-09-29; 修回日期: 2013-11-20

基金项目: 国家自然科学基金(51205185); 2012 年度江苏省“青蓝工程”中青年学术带头人项目

分类。

本文提出了使用随机森林算法进行特征选择的方法,以随机森林分类精度为准则函数,将每个特征对分类的贡献度量化,帮助我们了解在分类过程中哪些特征发挥了关键作用,同时通过这些特征定位到正常人和精神分裂症患者存在异常连接的功能脑区,作为临床诊断的客观对照。

2 fMRI 信号特征采集和特征提取

本文研究所使用的数据来源于哈佛大学,被试者为 11 名健康者和 11 名精神分裂症患者。22 名被试者参加了静息态 fMRI 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(TR)为 1 500 ms,回波时间(TE)为 27 ms,视野(FOV)为 24 cm×24 cm,翻转角(Flip Angle)为 90°,矩阵为 53×63,层厚 4 mm,层距 1 mm,共 46 层。从机器获得的初始文件为 DICOM 格式,并不能直接应用于数据分析,需要将其转化为可分析的文件格式,运用 SPM8 或相关软件将其转化为头文件(*.hdr)和图像文件(*.img)的格式。通过对时序信号序列的时间校正及空间校正消除机器、人体心跳呼吸等噪声数据,经过图像配准、平滑处理、去线性漂移、滤波后得到标准化的可分析数据格式。

得到上述预处理后的数据,通过和 AAL(Anatomical Automatic Labeling)标准模板匹配,每个脑图像的大脑被分成 90 个脑区^[6],通过计算 90 个脑区平均时间序列的皮尔森相关系数:

$$r_{ij} = \frac{\sum_{t=1}^T [x_i(t) - \bar{x}_i] \cdot [x_j(t) - \bar{x}_j]}{\sqrt{\sum_{t=1}^T [x_i(t) - \bar{x}_i]^2} \sqrt{\sum_{t=1}^T [x_j(t) - \bar{x}_j]^2}},$$

式中, x_i 和 x_j 分别表示体素 i 和体素 j 的时间序列, $\bar{x}_i$ 和 $\bar{x}_j$ 分别表示时间序列的均值,就可以得到脑区之间的功能连接系数矩阵。这是一个 90×90 的对称矩阵,可以提取上三角的矩阵元素展开得到一个 1×4 005 的矩阵,这就是获得的初始特征。

3 基于随机森林算法的特征选择

3.1 随机森林算法

随机森林算法^[7]是一种基于决策树的分类器算法,随机森林是由多个决策树 $\{h(x, \Theta_i), i = 1, 2, \dots, k\}$ 集成,其中 $\{\Theta_i\}$ 是独立同分布的随机变量。

随机森林有两个重要的随机性,训练集抽取的随机性和节点候选分割特征集合的随机性。这两个随机性保证了构成随机森林的决策树的多样性。在对一个输入向量中一个新的对象进行分类时,将输入向量放入到森林中的每棵树,每棵树给出一个分类,即每棵树对类进行“投票”,分类结果按树分类器的投票多少而定。整个算法过程包括两个部分:决策树的生长和投票过程。随机森林采用 Bagging 方法,应用自助法(bootstrap)^[8]重采样技术生成多个决策树分类器。

3.1.1 随机决策树的生成^[7]

(1)从容量为 N 的训练样本数据中随机的有放回的抽取 bootstrap 样本集,重复 k 次(即生成树的数目为 k),抽到的每个 bootstrap 样本集为决策树生长的训练集。每次未被抽到的样本组成了 k 个袋外数据(out of bag, OOB)。

(2)选取候选分割特征。设有 M 个输入特征,在决策树的每个节点处指定一个数 $m(m \leq M)$,从 M 个输入特征中随机选出 m 个特征,在那些 m 个特征中选择最佳的特征用于分割节点。然后重复上述步骤构造决策树各个分支,直到这棵树能够准确表示分类或者遍历所有属性。在森林的生长过程中, m 的值保持不变。

(3)随机森林的错误率取决于两个因素:森林中任意两棵树的相关性,森林中单棵树的强度。因此在决策树的生长过程中,不对其进行修剪,保证每棵决策树最大限度的生长,以使决策树具有高差异和低偏差特点。

3.1.2 随机森林投票法则

随机森林输出结果可以为每棵决策树输出结果的多种组合形式,最简单方式就是线性组合。被预测的样本数据,根据每棵决策树给出的预测类标,依据投票结果的多少,决定预测样本的最终归属类。投票

法则: $H(x) = \arg \max_Y \sum_{i=1}^k I(h_k(x) = Y)$, 其中 $H(x)$ 为组合分类模型, h_k 是单个决策树分类模型, Y 是目标变量, $I(\cdot)$ 为示性函数。

3.1.3 泛化误差的 OOB 估计^[9]

给定一组分类模型: $\{h_1(X), h_2(X), \dots, h_k(X)\}$, 每个分类数据集都是从原始数据集 (X, Y) 中随机抽样获得,余量函数(margin function)为 $mg(X, Y) = \max_k I(h_k(X) = Y) - \max_{j \neq k} I(h_j(X) = Y)$, 由余量函数可以引出泛化误差: $PE^* = P_{X,Y}(mg(X, Y) < 0)$, 当决策树分类器足够多,

$h_k(X) = h(X, \Theta_k)$ 服从强大数定律. 随着决策树数量的增加, 所有序列 $\Theta_1, \dots, \Theta_k, PE^*$ 几乎处处收敛于 $P_{X,Y}(P_{\Theta}(h(X, \Theta) = Y) - \max_{j \neq k} P_{\Theta}(h(X, \Theta) = j)) < 0$. 随机森林使用 Bagging 方法, 每棵决策树都可以得到一个 OOB 误差估计, 平均所有决策树的 OOB 误差估计, 即可得到随机森林的泛化误差.

3.2 fMRI 数据重要特征的选取

特征的选择过程其实对特征重要程度进行排序的过程. 在节点分割时, 以特征作为分割特征. 特征的好坏取决于与类标的相关度. 本文中选择的基尼指数作为衡量特征分割的效果. 假设集合 T 中包含 N 个不同类别记录的样品, 那么它的 Gini 指数为 $gini(T) = 1 - \sum_{j=1}^N P_j^2$, 其中 P_j 是第 j 类样本出现的概率. 如果在一次分割之后, 集合 T 被分成 m 部分, $N_1, N_2, \dots, N_m$. 那么这次分割的 $gini_{split}(T)$ 为 $gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \dots + \frac{N_m}{N} gini(T_m)$, 基尼指数大小反比与分割效果.

本文依据随机森林分类精度准则函数, 根据 GINI 指数对于重要特征的排名, 通过设定阈值, 选定贡献度排名阈值前的特征作为识别分类的主要特征. 挑选出特征子集后使用支持向量机(SVM)做分类识别不但可以缩短分类时间, 同时消除了冗余信息, 有利于提高识别精度. 而且由于数据自身的特点, 每个样本的 4 005 维数据即为 90 个脑区的功能连接相关系数矩阵, 可以根据特征排列的重要程度, 对照 AAL 模板识别出精神分裂症患者和健康者之间存在异常连接的脑区.

为了避免实验结果的不稳定性, 在特征选择中引进了 10 折交叉验证法. 即将原始数据集随机分成 10 等份, 其中的 9 份用于随机森林训练, 剩下的 1 份用于验证. 根据 GINI 指数, 挑选出 10 折交叉验证法中测试集分类精度最高的一次迭代产生的重要性排序特征作为特征子集. 算法主要流程见图 1.

4 实验和结果分析

为了验证提出方法的有效性, 使用静息态 fMRI 数据进行特征选择模式分类. 当设定生长 2 000 棵树, 且每个分裂节点设定的特征数为 63 个时, 实验效果较好. 图 2 所示为随机森林 10 折交叉验证法分类精度最高的特征贡献度结果, 图 2 中横坐标表示特征编号, 纵坐标分别表示平均下降分类误差值和平均下降 GINI 指标值, 其中这两个值越大表示

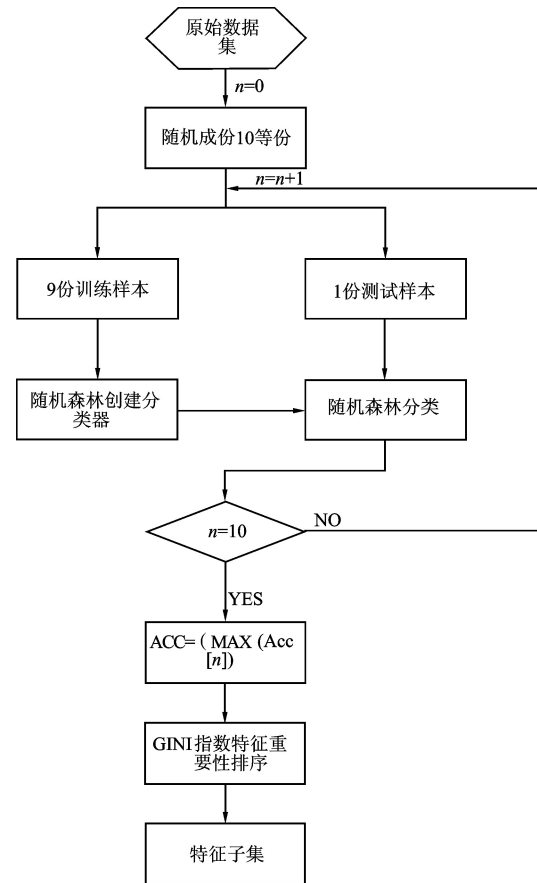


图 1 随机森林特征选择流程

(ACC 为分类精度, $Acc[n]$ 为 10 折交叉验证中各次分类精度)
特征在分类中作用越大, 即贡献度越大.

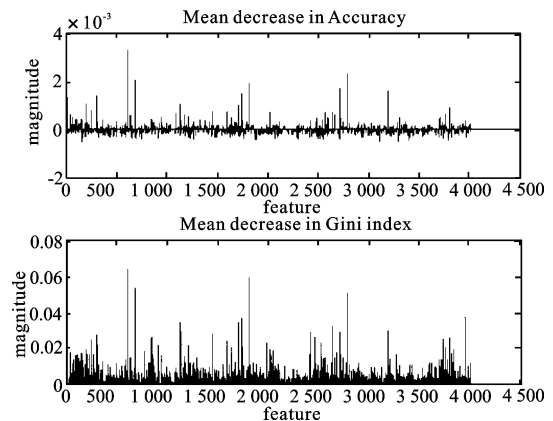


图 2 分类误差值和 GINI 指数值

为便于对比, 文中首先使用 PCA 技术对原始数据进行降维, 选择了集中 99% 能量的成分, 得到 18 个成分后使用 SVM 分类器选取前 12 个样本作为训练, 剩下的 10 个样本作为测试, 多次实验, 得到平均分类正确率为 70%; 使用随机森林算法对原始数据同样优选出 18 个特征后, 同样使用 SVM 分类器选取前 12 个样本作为训练, 剩下的 10 个样本作为

测试,多次实验,得到平均分类正确率为 85%。

由于样本数量的限制,为了检验分类器的泛化能力,采用留一法(Leave-One-Out Cross-Validation, LOOCV)交叉验证。对于分类结果有 3 个精度:分类精度,灵敏度,特异度,结果如表 1 所示。

表 1 分类结果的分类精度、灵敏度、特异度 %

	分类正确率	分类精度	灵敏度	特异度
PCA+SVM	70	59	60	58
RF+SVM	85	86.36	90	83.33

表 2 特征子集对应序号及对应 AAL 编号

序号	14	15	300	610	689	1126	1140	1451	1706	1743	1807	1809	2421	2637	2709	2781	3187	3952
AAL 编号	1	1	4	8	9	14	14	19	22	23	24	24	34	38	39	41	50	80
	15	16	40	16	14	61	75	21	69	39	37	39	46	48	69	42	52	82

通过 AAL 模板编号,可以找到对应的脑区,得出与文献[10]类似的精神分裂症患者异常脑区结果:背外侧前额叶皮层(AAL 编号 4),额中回(AAL 编号 8),海马(AAL 编号 37、38),海马旁回(AAL 编号 39、40),杏仁核(AAL 编号 41、42),颞横回(AAL 编号 80),颞上回(AAL 编号 82)。

5 结束语

本文运用结合了随机森林算法的特征选择方法对静息状态下的精神分裂症和健康者 fMRI 数据进行特征选择,并使用支持向量机对其进行分类,研究结果表明该算法可以对 fMRI 数据有效降维,在对重要特征的选取上表现出良好的性能,有效地提高了分类正确率,并且能根据特征的序号定位到相应脑区。实验过程中发现基于随机森林算法特征优选方法对抗噪声性能不够理想,对于使用不同样本会表现出一定的波动,这也是进一步研究的方向。

参考文献:

- [1] Brodersen KH, Schofield TM, Leff AP, et al. Generative embedding for model-based classification of fMRI data[J]. PLoS Comput Biol, 2011, 7(6):1-19.
- [2] 吴晓婷,闫德勤. 数据降维方法分析与研究[J]. 计算机应用研究, 2009, 26(8):2832-2835.
- [3] 黎银环,钟艳花. 入侵检测中的混合特征选择算法研究[J]. 微电子学与计算机, 2012, 29(9):51-54.
- [4] Shen H, Wang LB, Liu YD, et al. Discriminative analysis of resting-state functional connectivity patterns

从表 1 中可以明显看出运用随机森林进行特征优选后用支持向量机分类,可以使分类精度达到 85%,高于运用 PCA 降维后的支持向量机分类,同时通过留一交叉验证法对样本数据的分类,分类结果不管是分类精度,灵敏度或特异度都全面高于结合 PCA 降维的支持向量机分类器,说明了结合随机森林特征优选的支持向量机分类器不仅分类正确率高,并且分类器泛化能力更强。

根据上述实验结果得到 18 个重要特征的序号,由特征序号得到对应的 AAL 编号如表 2 所示。

- of schizophrenia using low dimensional embedding of fMRI[J]. Neuroimage, 2010, 49(4):3110-3121.
- [5] 曹芳. 精神分裂症在静息态功能性磁共振成像下的多模式分析[D]. 长沙:中南大学, 2012.
 - [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] Breiman L. Random forests[J]. Machine learning, 2001, 45(1):5-32.
 - [8] Efron B, Tibshirani R J. Bootstrap measures for standard error confidence interval and other measures of statistical accuracy[J]. Statistical Science, 1986, 1(1):54-74.
 - [9] Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: a survey and results of new tests[J]. Pattern Recognition, 2011(44):330-349.
 - [10] Archana Venkataraman, Thomas J. Whitford, et al. Whole brain resting state functional connectivity abnormalities in schizophrenia[J]. Schizophrenia Research, 2012, 139(1/3):7-12.

作者简介:

袁晓龙 男, (1989-), 硕士研究生. 研究方向为医学图像处理, 模式识别等。

梅雪 女, (1975-), 博士, 副教授. 研究方向为模式识别与计算机视觉等。