

KPCA和Adaboost算法在阿尔茨海默症功能磁共振影像分类中的应用

李长胜,王瑜,肖洪兵,邢素霞

北京工商大学计算机与信息工程学院食品安全大数据技术北京市重点实验室,北京 100048

【摘要】本研究的目的在于使用机器学习方法,对脑部功能磁共振成像数据进行分析与特征提取,完成对阿尔茨海默症(AD)的辅助诊断与分析。首先对数据进行预处理与去除协变量,并从大脑全局特征出发,根据现有的自动解剖标记模板,把每个被试的大脑分为116个脑区,通过提取每个脑区的时间序列,构建全脑功能连接矩阵,然后使用核主成分分析法进行特征提取,最后用Adaboost算法进行分类。在对34名AD患者、35名轻度认知障碍患者和35名正常对照组的功能磁共振成像数据进行的实验结果表明,利用静息态功能磁共振成像,同时结合机器学习的方法,能够有效地实现AD的正确分类,准确率可以达到96%,该结果可以为AD患者的临床辅助诊断提供有效的判断依据。

【关键词】功能磁共振成像;阿尔茨海默症;轻度认知障碍;功能连接矩阵;核主成分分析

【中图分类号】R445.2;R318

【文献标志码】A

【文章编号】1005-202X(2019)07-0784-05

Application of KPCA and Adaboost algorithm in the classification of functional magnetic resonance images of Alzheimer's disease

LI Changsheng, WANG Yu, XIAO Hongbing, XING Suxia

Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China

Abstract: The purpose of this study is to achieve the auxiliary diagnosis and analysis of Alzheimer's disease (AD) by analyzing and characterizing brain functional magnetic resonance imaging (fMRI) data using machine learning method. After the fMRI data is preprocessed and the covariate is removed, the brain of each subject is divided into 116 brain regions according to anatomical automatic labeling template, and the whole brain functional connection matrix is constructed by extracting the time series of each brain region. Kernel principal component analysis is used to extract features and Adaboost algorithm is used for classification. The results of the experiment on fMRI images of 34 patients with AD, 35 patients with mild cognitive impairments and 35 normal controls show that using resting state fMRI combined with machine learning method can effectively realize the accurate classification of AD, with a classification accuracy rate up to 96%. The proposed method can provide an effective basis for the auxiliary diagnosis of patients with AD.

Keywords: functional magnetic resonance imaging; Alzheimer's disease; mild cognitive impairment; functional connection matrix; kernel principal component analysis

前言

阿尔茨海默症(Alzheimer's Disease, AD)是老年人

中最常见的疾病之一,AD患者数量的增长不断加剧家庭和社会的负担,在早期阶段开展该疾病的预防和治疗是最有效的^[1]。因此,AD的早期诊断意义重大。传统的AD临床诊断方法过程复杂,取决于临床病史和临床评估,检查周期长^[2]。神经影像学可以测量AD患者脑部的病理性变化,在过去的十年中,这些措施已经越来越多地成为诊断和预测的工具,一个有效的计算机自动诊断系统可以很好地起到疾病辅助诊断的作用^[3]。

伴随着神经影像学技术的飞速发展,核磁共振成像(MRI)技术作为一种新兴的医学成像技术在临床医

【收稿日期】2019-03-07

【基金项目】国家自然科学基金(61671028);北京市自然科学基金面上项目(4162018);国家重大科技研发子课题(ZLJC6 03-5-1);北京工商大学校级两科培育基金(19008001270)

【作者简介】李长胜,硕士研究生,研究方向:计算机视觉、医学图像处理、模式识别,E-mail: 516795305@qq.com

【通信作者】王瑜,博士,副教授,研究方向:计算机视觉、医学图像处理、模式识别,E-mail: wangyu@btbu.edu.cn

疗方面得到了广泛的应用,通过结合机器学习的方法,对MRI数据进行分析,从而达到辅助医生对AD患者进行诊断的目的。其中结构磁共振成像(sMRI)数据较好地反应了脑组织的形态学特征^[4],而静息态功能磁共振成像(fMRI)反映了大脑中各脑区的激活状态^[5]。研究表明,AD患者的大脑功能连接网络与正常人有明显的差异^[6]。2016年,Guo等^[7]通过实验证明小脑的萎缩与AD等常见神经疾病的关系。因此,本研究通过对被试的fMRI图像数据提取包括小脑在内的全脑功能连接矩阵作为初始特征,使用核主成分分析(Kernel Principal Component Analysis, KPCA)方法降维,并结合机器学习方法,实现对AD患者、轻度认知障碍(Mild Cognitive Impairment, MCI)患者和正常对照组(Normal Control, NC)的分类,从而准确地对被试的患病情况进行判别,以达到辅助诊断的效果。

1 数据及预处理

1.1 数据介绍

本实验所收集的静息态fMRI数据,全部来源于AD神经影像学(Alzheimer's Disease Neuroimaging Initiative, ADNI)数据库^[8],实验共选取104例被试的静息态fMRI图像数据,其中包括34名AD患者、35名MCI患者和35名NC,每个被试分别有140幅图像,每幅图像扫描48层,所有被试样本的年龄、性别信息如表1所示。

表1 被试者统计分析
Tab.1 Clinical information of subjects

组别	<i>n</i>	男/女	年龄/岁
AD	34	18/16	73.29±7.65
MCI	35	13/22	73.34±8.43
NC	35	20/15	77.11±6.69

AD:阿尔茨海默症;MCI:轻度认知障碍;NC:正常对照

1.2 数据预处理

本实验主要使用Statistical Parametric Mapping (SPM 8)和RS-fMRI Data Analysis Toolkit (REST)两个软件包在MATLAB平台对获取的fMRI图像数据进行预处理操作^[9]。实验环境为个人PC机,处理器: Intel®Corei5-4200 H, CPU@2.80 GHz,内存为4 GB,实验运行环境为MATLAB2017a。从ANDI获取的数据格式为fMRI常用的Analyze格式,对每个被试的静息态fMRI图像数据依次进行如下处理:首先去除前10个时间点的图像数据;以第47层作为参考层进行

时间层校正;进行头动校正;对图像进行空间标准化;对图像进行空间平滑(高斯核半宽全高设为6 mm×6 mm×6 mm);对平滑后的图像进行去线性漂移和0.01~0.08 Hz的低频滤波;最后去除头动校正时生成的头动参数的协变量,去除全脑均值信号、白质信号和脑脊液信号的协变量^[10]。

2 特征提取与分类器的选择

2.1 功能连接矩阵

所有被试的静息态fMRI图像进行预处理之后,根据自动解剖标记(Anatomical Automatic Labeling, AAL)分区模板^[11],每个被试的脑部MRI被分成116个脑区。其中,被试的大脑部分共被分为90个脑区,小脑部分被分为26个脑区,AAL模板脑区分割情况如图1所示。

按照AAL模板,先将脑部MRI划分为116个脑区,接下来将每个被试对应的经预处理后得到的fMRI图像,结合AAL模板划分的脑区情况,分别计算划分后每个脑区内的体素平均值,从而组成脑区内体素平均值变化的时间序列,因此每个被试都能获取116个脑区的平均时间序列,数学形式即为一个116×130的矩阵,最后通过计算皮尔逊相关系数,得到一个全脑功能连接矩阵^[12]。功能连接矩阵中元素的含义对应表示相应位置的脑区之间的功能连接情况,其中如果值大于零则表示功能连接正相关,值小于零表示功能连接负相关。由于通过计算皮尔逊相关系数的过程可以得到一个对称的功能连接矩阵,因此只需取矩阵中下三角的数据作为用于机器学习分类的初始特征,如图2所示,获取功能连接矩阵流程如图3所示^[13]。

2.2 KPCA

利用上述过程,可以得到功能连接矩阵,但由于其本身维度较大,需要对其进行降维,从而更好地利用机器学习的方法进行分类。Pearson^[14]提出的主成分分析(Principal Component Analysis, PCA)方法是常用的降维方法之一,PCA不仅可以实现对高维数据的降维,更重要的是可以通过降维去除原本数据中的噪声,发现数据中的模式。PCA用相对更少的*m*维特征取代原始数据中的*n*维特征,最终得到的新特征是原始特征的线性组合,这些组合最终使样本方差最大化,达到实现新的*m*维特征互不相关的目的。由于本实验中选取大脑功能连接矩阵作为初始特征,特征之间的关系并不是线性的,而常用的PCA方法是一种线性降维方法,无法挖掘特征之间的非线性关系,因此影响了PCA的效果。

KPCA通过使用核函数的方法^[15],将样本的空间通过核函数映射到更高维的空间中,再利用这个高维空间进行线性降维,从而很好地解决了非线性特



图1 AAL分区模板

Fig.1 Anatomical automatic labeling template

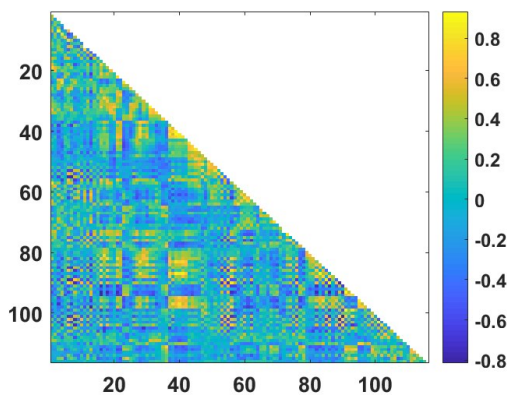


图2 功能连接矩阵

Fig.2 Functional connection matrix

征的降维问题。假设在高维空间中将一组数据 $X=(x_1, x_2, \dots, x_n)^T$ 投影到由 $W=(w_1, w_2, \dots, w_d)$ 确定的超平面上,根据PCA方法所得出的结论^[16]为:

$$XX^T w_i = \lambda_i w_i \tag{1}$$

设 z_i 为样本点 x_i 在高维特征空间的像,则对于 w_j 有:

$$\left(\sum_{i=1}^m z_i z_i^T\right) w_j = \lambda_j w_j \tag{2}$$

所以得到:

$$w_j = \sum_{i=1}^m z_i \alpha_i^j \tag{3}$$

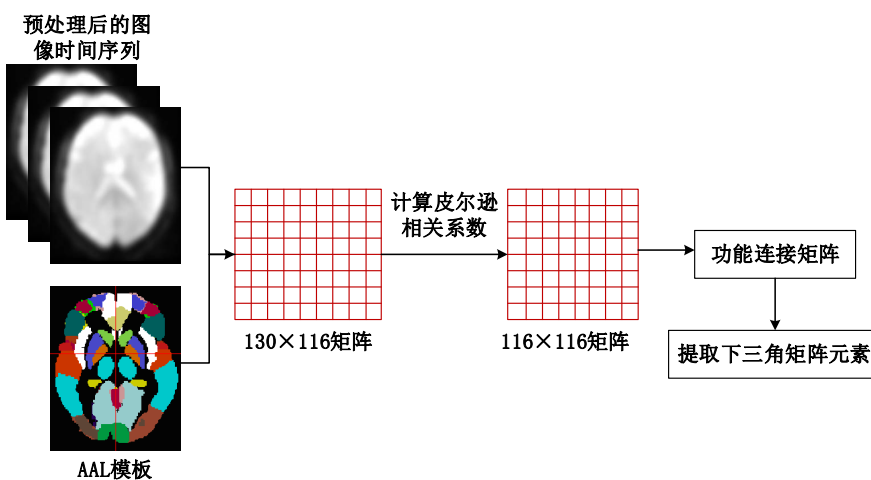


图3 获取功能连接矩阵流程

Fig.3 Acquisition process of functional connection matrix

其中, $\alpha_i^j = \frac{1}{\lambda_j} z_i^T w_j$ 是 α_i 的第 j 个分量。

假设 z_i 是由原始数据中属性空间的样本点 x_i 通过映射关系 ϕ 产生,即 $z_i = \phi(x_i)$, $i = 1, 2, \dots, m$,若 ϕ 能被显式表达出来,则通过它将样本映射至高维特征空间,再在特征空间实施PCA即可。则式(2)变换为:

$$\left(\sum_{i=1}^m \phi(x_i) \phi(x_i)^T\right) w_j = \lambda_j w_j \tag{4}$$

式(3)变换为:

$$w_j = \sum_{i=1}^m \phi(x_i) \alpha_i^j \tag{5}$$

由于映射关系 ϕ 的具体形式一般情况下很难寻找,因此引入核函数 $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$,有:

$$K \alpha^j = \lambda_j \alpha^j \tag{6}$$

其中, K 为 k 对应的核矩阵, $(K)_{ij} = k(x_i, x_j)$,

$\alpha^j = (\alpha_1^j; \alpha_2^j; \dots; \alpha_m^j)$ 。式(6)是特征分解的问题,因此可以得到 K 最大的 d 个特征值对应的特征向量。

对新样本 x , 其投影后的第 $j(j=1, 2, \dots, d)$ 维坐标为:

$$z_j = w_j^T \phi(x) = \sum_{i=1}^m \alpha_i^j \phi(x_i)^T \phi(x) = \sum_{i=1}^m \alpha_i^j k(x_i, x) \quad (7)$$

KPCA 虽然在求解过程中与 PCA 有类似的地方,但是使用 KPCA 方法降维与 PCA 方法具有本质上的差别,因为 KPCA 引入的非线性映射函数,间接去掉了原样本中的各个属性之间的非线性相关性,最终通过 PCA 方法实现降维目的,最大程度保留了样本的主要特征信息,同时简化了数据。

2.3 分类器

本研究选用目前比较经典的两种分类方法进行实验,包括支持向量机(Support Vector Machine, SVM)和 Adaboost 方法。

2.3.1 SVM SVM 是机器学习方法中一种常见的有监督的学习模型^[17], SVM 根据支持向量构造最优超平面,使得数据集中所有向量都能够满足被某一个超平面正确分类,且距离该超平面的支持向量与超平面之间的距离最大,即分类后的不同组数据之间距离间隔最大。

假设数据样本为 x_1, x_2, \dots, x_n , 在训练样本中,模型的支持向量和距离其最近的平行超平面的数学公式可以表达为:

$$\begin{cases} w^T x - b = 1 \\ w^T x - b = -1 \end{cases} \quad (8)$$

其中, w 为分类超平面的法向量; x 为分类超平面上的点; b 为位移量,两个分类超平面之间具有最大间隔,且他们之间的距离为 $\frac{2}{\|w\|^2}$, 因此只要使 $\|w\|^2$ 最小化,就可以使这两个超平面之间的距离达到最大化, SVM 的目标函数^[18]为:

$$\min \|w\| \text{s.t. } y_i(w^T x_i + b) \geq 1, i = 1, \dots, n \quad (9)$$

找到使 SVM 目标函数最小的 w 和 b , 也就找到了最合适的超平面,实现对数据的分类。

2.3.2 Adaboost Adaboost 是一种较为先进的模式识别分类算法^[19], 可以通过不断调整数据集中每个样本对应的权重,实现分类学习的目的。首先在每个样本对应初始权重时训练出一个弱分类器,然后根据分类的结果,重点关注被分错的样本,逐渐增加未被正确分类样本的权重值,同时降低被正确分类样本的权重值,然后在新的样本分布下再次对弱分类器进行训练。以此类推,经过 T 次循环训练之后,得到 T 个弱分类器。最后赋予训练好的 T 个弱分类器一定的权重并叠加,最终通过组合得到一个强分类器。

给定训练样本以及对应的分类 $(x_1, y_1), \dots, (x_n, y_n)$,

其中 $x_i \in X, y_i \in Y = \{-1, 1\}$ 。初始化样本权重为 $D_1(i) = \frac{1}{n}$, 即训练样本的初始权值分布,然后进行最大循环次数为 T 的迭代训练。Adaboost 算法流程如下:(1)对于第 $t=1, \dots, T$ 次迭代,通过使用样本的权值分布 D_t 训练弱分类器;(2)得到弱分类器的分类结果,计算如下弱分类器分类结果的错误率:

$$\varepsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i] \quad (10)$$

(3)选取参数 α_t :

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1 - \varepsilon_t}{\varepsilon_t}\right) \quad (11)$$

更新第 $t+1$ 次迭代中的权值分布:

$$\begin{aligned} D_{t+1}(i) &= \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t}, & \text{若 } h_t(x_i) = y_i \\ e^{\alpha_t}, & \text{若 } h_t(x_i) \neq y_i \end{cases} \\ &= \frac{D_t(i)}{Z_t} \exp\{-\alpha_t y_i h_t(x_i)\} \end{aligned} \quad (12)$$

其中, Z_t 是归一化参数;(4)在经过 T 次迭代训练之后,得到 h_1, h_2, \dots, h_T , 共 T 个弱分类器;(5)最后通过加权投票,组合得到强分类器:

$$H(x) = \text{sign}\left(\sum_{i=1}^T \alpha_i h_i(x)\right) \quad (13)$$

本研究采用 Scikit-learn 中实现的 Adaboost 方法^[20], 其中选择的弱分类器使用分类回归树方法。

3 实验结果与分析

为证明本研究提出方法的有效性,精心设计两组实验。本实验根据采集的 fMRI 图像数据,利用机器学习的方法实现 AD、MCI 和 NC 的分类,根据每一个被试的 fMRI 图像分别计算功能连接矩阵,然后对功能连接矩阵进行 KPCA 降维后作为特征,得到特征向量。随机选取所有样本的 70% 作为训练集,剩余 30% 作为测试集,最后根据得到的特征向量,通过 SVM 和 Adaboost 两种分类器进行分类,观察分类效果。

实验选取的评价指标为准确率(precision)、召回率(recall)和 F1 值(F1 score),计算公式如下:

$$\text{precision} = \frac{TP}{TP + FP} \quad (14)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (15)$$

$$\text{F1 score} = \frac{2PR}{P + R} \quad (16)$$

其中, P 和 R 分别代表准确率和召回率, TP(True Positive) 为分类后被正确标记为正例的个数; FP(False Positive) 为分类后被错误标记为正例的个数; TN(True Negative) 为分类后被正确标记为负例的个数; FN(False Negative) 为分类后被错误标记为负例的个数。在结果的分析过程中,精确率和召回率的含义也可以被理解为查准率和查全率,而 F1 值可以

被看作是模型准确率和召回率的一种加权平均,作为统一准确率和召回率的评估标准,来衡量模型分类的性能。具体实验结果如表2所示。

根据表2的结果可以看出,通过使用KPCA降维后的特征,并结合Adaboost方法,最终分类效果要优于使用传统的SVM分类器。并且KPCA结合Adaboost方法分类准确率最高可达到96%,F1值最高0.95,说明全脑功能连接矩阵可以作为区分AD、NC和MCI的特征,同时KPCA可以有效地从数据中继续提取出更多的非线性特征,通过Adaboost分类算法实现有效区分AD、NC和MCI,以达到辅助诊断的目的。

表2 实验对比结果

Tab.2 Comparison of experimental results

方法	评价指标	AD vs NC	NC vs MCI	AD vs MCI
PCA+SVM	准确率/%	53	56	57
	召回率/%	54	52	55
	F1值	0.53	0.48	0.55
KPCA+SVM	准确率/%	61	65	61
	召回率/%	67	55	62
	F1值	0.63	0.50	0.56
PCA+Adaboost	准确率/%	60	57	51
	召回率/%	57	57	52
	F1值	0.55	0.57	0.47
KPCA+Adaboost	准确率/%	92	96	78
	召回率/%	90	95	76
	F1值	0.90	0.95	0.76

PCA:主成分分析;SVM:支持向量机;KPCA:核主成分分析

4 总结

本研究对被试的fMRI数据计算其包括小脑在内的全脑功能连接矩阵,利用KPCA降维并作为特征,然后分别采用SVM和Adaboost方法对AD、MCI和NC进行分类。实验结果表明,本研究提出的方法可以对fMRI数据进行有效分析,由于KPCA方法能够挖掘数据的非线性信息,所以KPCA方法较PCA方法获得了更好的准确率,同SVM方法相比,Adaboost分类器取得了更好的分类效果,同时也说明全脑功能连接矩阵能够反映出AD、MCI和NC之间的差异,这些结论可以为AD的诊断提供有效的判断依据。

【参考文献】

[1] ZHANG J, GAO Y, GAO Y Z, et al. Detecting anatomical landmarks

for fast Alzheimer's disease diagnosis[J]. IEEE Trans Med Imaging, 2016, 35(12): 2524-2533.

- [2] 樊东琼,李锐,雷旭,等.阿尔兹海默症及轻度认知障碍静息态大尺度脑网络功能连接的变化[J].心理科学进展,2016,24(2):217-227. FAN D Q, LI R, LEI X, et al. Changes in functional connectivity of large-scale brain networks in resting state of Alzheimer's disease and mild cognitive impairment [J]. Advances in Psychological Science, 2016, 24(2): 217-227.
- [3] RATHORE S, HABES M, AKSAM I M, et al. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages [J]. Neuroimage, 2017, 155: 530.
- [4] 周文,王瑜,肖红兵,等.基于KPCA算法的阿尔兹海默症辅助诊断[J].中国医学物理学杂志,2018,35(4):404-409. ZHOU W, WANG Y, XIAO H B, et al. Assisted diagnosis of Alzheimer's disease based on KPCA algorithm [J]. Chinese Journal of Medical Physics, 2018, 35(4): 404-409.
- [5] KHAZAEI A, EBRAHIMZADEH A, BABAJANIFEREMI A. Classification of patients with MCI and AD from healthy controls using directed graph measures of resting-state fMRI[J]. Behav Brain Res, 2017, 322(Pt B): 339-350.
- [6] FORD J, FARID H, MAKEDON F, et al. Patient classification of fMRI activation maps [C]// Medical Image Computing and Computer-Assisted Intervention. DBLP, 2003: 58-65.
- [7] GUO C C, TAN R, HODGES J R, et al. Network-selective vulnerability of the human cerebellum to Alzheimer's disease and frontotemporal dementia[J]. Brain, 2016, 139(Pt 5): 1527-1538.
- [8] WEINER M W, VEITCH D P, AISEN P S, et al. Recent publications from the Alzheimer's disease neuroimaging initiative: reviewing progress toward improved AD clinical trials[J]. Alzheimers Dement, 2017, 13(4): e1.
- [9] SONG X W, DONG Z Y, LONG X Y, et al. REST: a toolkit for resting-state functional magnetic resonance imaging data processing[J]. PLoS One, 2011, 6(9): e25031.
- [10] KWAK Y, PELTIER S J, BOHNNEN N I, et al. L-DOPA changes spontaneous low-frequency BOLD signal oscillations in Parkinson's disease: a resting state fMRI study[J]. Front Syst Neurosci, 2012, 6(6): 52.
- [11] 李亚鹏,覃媛媛,李炜.阿尔兹海默病患者大脑功能网络的变化[J].中国医学物理学杂志,2013,30(6):4510-4514. LI Y P, QIN Y Y, LI W. The functional brain network changes of Alzheimer's disease[J]. Chinese Journal of Medical Physics, 2013, 30(6): 4510-4514.
- [12] ZHANG M, XIA Z F, ZHANG F, et al. Cognitive effects of high-frequency repetitive transcranial magnetic stimulation in Alzheimer's disease: a pilot clinical study[J]. Alzheimers Dement, 2017, 13(7): P260-P261.
- [13] 刘美洁.脑磁共振成像数据的多类模式分析[D].长沙:国防科学技术大学,2011. LIU M J. Multi-class pattern analysis on human brain MRI dataset [D]. Changsha: National University of Defense Technology, 2011.
- [14] PEARSON K. On lines and planes of closest fit to systems of points in space[J]. Philos Mag, 1901, 2(6): 559-572.
- [15] SCHÖLKOPF B, SMOLA A, MÜLLER K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Comput, 1996, 10(5): 1299-1319.
- [16] ZHANG J, TUO X, YUAN Z, et al. Analysis of fMRI data using an integrated principal component analysis and supervised affinity propagation clustering approach[J]. IEEE Trans Biomed Eng, 2011, 58(11): 3184-3196.
- [17] CORTES C, VAPNIK V. Support-vector networks[J]. Mach Learn, 1995, 20(3): 273-297.
- [18] WANG Z, CHILDRESS A R, WANG J, et al. Support vector machine learning-based fMRI data group analysis[J]. Neuroimage, 2007, 36(4): 1139-1151.
- [19] ZHU J, ZOU H, ROSSET S, et al. Multi-class AdaBoost[J]. Stat Interface, 2009, 2(3): 349-360.
- [20] PEDREGOSA F, GRAMFORT A, MICHEL V, et al. Scikit-learn: machine learning in python[J]. J Mach Learn Res, 2013, 12(10): 2825-2830.

(编辑:陈丽霞)