·人工智能及识别技术· 文章编号:1000-3428(2016)12-0151-05 文献标志码:A 中图分类号:TP751.1

基于相似度矩阵约减的仿射聚类 fMRI 数据分析

管秀英,曾卫明,王倪传

(上海海事大学 信息工程学院,上海 201306)

摘 要: 利用仿射聚类(APC)方法分析数据量庞大的功能磁共振成像(fMRI)数据时,在时间复杂度、数据存储和聚类效果等方面存在局限性。为此,提出一种融合稀疏仿射传播聚类(SAPC)和相似度矩阵约减的新方法(SDAPC)。对 fMRI 数据进行稀疏逼近后,结合高斯密度函数和欧式距离对稀疏数据进行密度分析,完成约减后 fMRI 数据的功能连通性检测。任务态数据实验结果表明,对于单被试,SDAPC 的 ROC 曲线与 SAPC 接近,但运行速度比 SAPC 提高了约 3 倍;对于多被试,SDAPC 和 SAPC 的 ROC 曲线效果均优于其单被试的 ROC 曲线。静息态数据实验结果进一步表明,SDAPC 能成功提取出 9 个静息态脑网络。

关键词: 仿射传播聚类;功能磁共振成像;时间复杂度;相似度矩阵约减;高斯密度函数

中文引用格式: 管秀英,曾卫明,王倪传. 基于相似度矩阵约减的仿射聚类 fMRI 数据分析[J]. 计算机工程,2016,42(12):151-155.

英文引用格式: Guan Xiuying, Zeng Weiming, Wang Nizhuan. fMRI Data Analysis of Affinity Propagation Clustering Based on Similarity Matrix Reduction[J]. Computer Engineering, 2016, 42(12):151-155.

fMRI Data Analysis of Affinity Propagation Clustering Based on Similarity Matrix Reduction

GUAN Xiuying, ZENG Weiming, WANG Nizhuan

(College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

【Abstract】 Affinity Propagation Clustering (APC) method shows its limitations in time complexity, data storage and clustering results while handling massive functional Magnetic Resonance Imaging (fMRI) data. Aiming at these problems, this paper proposes a new method named SDAPC, which combines Sparse APC (SAPC) with similarity matrix reduction. It starts from sparse approximation on fMRI data, continues with the density analysis on sparse data by Gaussian density function and Euclidean distance, and finally realizes the detection on the functional connectivity of reduced fMRI data. The task-related data experiment gets the following results: SDAPC produces a fine ROC curve for single subject while running about three times faster than SAPC. SDAPC and SAPC both get better ROC curves for multiple subjects than single subject. The resting-state data experiment leads to the further finding that SDAPC can successfully identify nine resting-state networks.

【Key words】 Affinity Propagation Clustering (APC); functional Magnetic Resonance Imaging (fMRI); time complexity; similarity matrix reduction; Gaussian density function

DOI: 10.3969/j.issn.1000-3428.2016.12.027

0 概述

基于血氧水平依赖效应的功能磁共振成(functional Magnetic Resonance Imaging, fMRI)技术能有效地进行脑区域连通性检测。仿射传播聚类(Affinity Propagation Clustering, APC)[1]最先由 Frey 和 Dueck 教授提出,目前已应用到 fMRI 数据分析并得到了较好的结果[2-3]。APC 假设所有的数据点为初始类心,然后基于数据点之间的信息传递来确定最终的类心集。APC 需要计算相似度矩阵、吸引度矩阵和归属度矩阵,且在每次迭代过程中还要不断更新吸引度矩阵和归属度矩阵,这就导致数据空间存储大、计算复杂度高和算法运行时间长的问题,降低了算法的运行效率,并且 APC 对大于 3 000 的数据量难以聚出好的效果[4]。文献[5]提出了稀疏仿射传播聚类(Sparse APC, SAPC)算法,先对 fMRI 数据进行稀疏逼近,再

基金项目: 国家自然科学基金(31170952,31470954);上海市教育委员会科研创新重点项目(11ZZ143)。

作者简介: 管秀英(1991—),女,硕士研究生,主研方向为模式识别、图像处理;曾卫明,教授、博士、博士生导师;王倪传,博士研究生。

收稿日期: 2015-12-04 **修回日期:** 2016-01-25 **E-mail:** 2039049351@qq.com

进行仿射传播聚类,不仅降低了原始 APC 的运行时间,而且能运用到静息态数据分析中。但对于静息态的 fMRI 数据,通常稀疏逼近后的数据量达到 4 万多,SAPC 需要运行 40 min 左右[5]。

针对上述问题,本文提出一种基于稀疏变换[5-7]和数据点密度[8-9]的仿射传播聚类算法(SDAPC)。传统的 APC 是基于欧氏距离分析,但欧氏距离仅反映数据点的局部分布[9],忽略了数据的整体分布。因此,本文引入数据点密度概念,以同时反映数据点的局部分布和整体分布[8-9]。由于密度小的点很可能为噪声点,因此将之剔除以实现数据去噪。在单被试的基础上还进行多被试 fMRI 数据分析,考虑到文献[5]直接对多组被试数据求平均的方法具有一定局限性,因此,提出被试间聚类的方法进行实验,克服直接求平均带来的弊端,并且验证 APC 算法的有效性。

1 模型和方法

SDAPC 模型主要包含以下过程:先基于小波包分解获取稀疏逼近系数,再基于密度大小原则对相似度矩阵约减,最后对约减数据进行仿射传播聚类并基于最优类心集进行脑图谱重建。

1.1 基于小波包的稀疏逼近

脑功能网络检测可建模成盲源分离问题[10]。假定大脑中潜在的源信号为 $\boldsymbol{S}=(s_1\ s_2\cdots s_N)^{\mathrm{T}}\in\mathbb{R}^N$,fMRI 混合信号为 $\boldsymbol{X}=(x_1\ x_2\cdots x_P)^{\mathrm{T}}\in\mathbb{R}^P$,则盲源分离问题可表示为:

$$\boldsymbol{X}=\boldsymbol{AS} \tag{1}$$

其中,$\boldsymbol{A}$ 是混合矩阵,大小为 $P\times N$;行向量 $\boldsymbol{s}_i$ 和 $\boldsymbol{x}_i$ 的大小均为 $1\times M$,M 为大脑空间图谱点的个数。根据稀疏逼近理论[6-7],源信号 $\boldsymbol{S}$ 和混合信号 $\boldsymbol{X}$ 的稀疏表达形式分别为:

$$\begin{cases}\boldsymbol{S}=C_S\boldsymbol{\Phi}\\ \boldsymbol{X}\approx C_X\boldsymbol{\Phi}\end{cases} \tag{2}$$

其中,C_X 是信号 $\boldsymbol{X}$ 和 $\boldsymbol{S}$ 在完备字典 $\boldsymbol{\Phi}$ 上的稀疏表示系数。结合式(1)和式(2)可得:

$$C_x\approx\boldsymbol{A}C_S \tag{3}$$

文献[11-14]研究表明,可通过小波包分解有效地获取混合信号 X 的稀疏逼近系数 C_X。

1.2 基于密度的相似度矩阵约减

经小波包得到的 fMRI 数据量依然非常庞大,APC 运行时间过长[5],因此,要对稀疏逼近系数 C_X 做进一步约减。

设体素点个数为 Q,则体素点集合为 $C_x=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_Q\}$,其中,$\boldsymbol{x}_i$ 是基于时间的一维向量,大小为 $1\times T$。本文采用欧式距离计算体素点的相似性,利用基于欧式距离的高斯密度函数计算数据点的密度大小[8-9]。

在统计学与概率论中,高斯函数是正态分布的密度函数,根据中心极限定理,它是复杂总和的有限概率分布。高斯密度函数的数学表达式为:

$$f(x)=\frac{1}{\sqrt{2\pi}\delta}\mathrm{e}^{-\frac{(x-\mu)^2}{2\delta^2}} \tag{4}$$

其中,μ 为均值;δ 为方差。当 $\mu=0,\delta=1$ 时,式(4)变为标准高斯密度函数,形式如下:

$$f(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}} \tag{5}$$

研究表明,fMRI 数据存在着自然分布[8],有的数据点周围的数据点多,则该点密度大;有的数据点周围的数据点少,则该点密度小。密度小的点可能为噪声点,将之剔除。基于欧式距离的高斯密度函数构造形式如下:

$$f(\boldsymbol{x}_i,\boldsymbol{x}_j)=\mathrm{e}^{-\frac{d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)}{2\delta^2}} \tag{6}$$

其中,δ 为密度参数,$\delta=1$。欧式距离非负,所以,高斯密度函数在 x 正半轴有值。因为高斯密度函数在 x 正半轴单调递减,所以当两数据点的欧式距离越大,则函数密度值越小。$\boldsymbol{x}_i$ 数据点的密度大小为:

$$density(\boldsymbol{x}_i)=\sum_{j=1}^{Q}\mathrm{e}^{-\frac{d(\boldsymbol{x}_i,\boldsymbol{x}_j)^2}{2\delta^2}} \tag{7}$$

其中,Q 为体素点的个数。将 x_i 的密度大小与所有数据点的密度大小之和的比值称为密度权重,密度权重刻画了 x_i 数据点对所有数据点影响的贡献率,表示如下:

$$W_{\mathrm{density}}(\boldsymbol{x}_i)=\frac{density(\boldsymbol{x}_i)}{\sum_{i=1}^{Q}density(\boldsymbol{x}_i)} \tag{8}$$

其中,x_i 数据点的密度权重越大,说明 x_i 对所有数据点的影响越大,反之越小。密度权重小的点(本文取 $W_{\mathrm{density}}(\boldsymbol{x}_i)<0.01$)很可能为噪声点,将之剔除。将密度权重大的数据点挑出,计算它们之间的相似度,作为 APC 算法的输入矩阵。

1.3 仿射传播聚类算法

APC 算法先输入相似系数矩阵,再基于数据点之间的信息传递不断更新算法,最后基于能量函数 $E(c)$[1]最大得出最优类心集。

本文借鉴文献[5]的分组思想,提出三次聚类方法。假设稀疏约减后的数据点个数为 q,对 q 个数据进行分组,设每组数据量为 y,则组数为 $K=\lceil q/y\rceil$,其中$\lceil\ \rceil$为上取整记号。三次聚类思想如下:

1)分别对每组数据进行一次 APC 聚类,得到 k 组类心集。

2)把 k 组类心集组合起来,进行二次 APC 聚类(单被试聚类),得到单被试的类心集。

3)采集 6 个被试的 fMRI 数据,对每个被试都进行上述的稀疏、约减和二次 APC 聚类,最终得到6 组类心集。组合 6 组类心集,进行三次 APC 聚类(多被试聚类),得到最终的聚类结果,即式(1)中的混合矩阵 $\boldsymbol{A}$。

对混合矩阵 $\boldsymbol{A}$ 取逆得到解混矩阵 $\boldsymbol{W}$。源信号 $\boldsymbol{S}$ 可由下面公式得到:

$$\boldsymbol{S}=\boldsymbol{XW} \tag{9}$$

2　实验数据处理

fMRI 数据的预处理软件是 Matlab 的 spm8。fMRI 数据的 GLM 模板生成工具包是 GIFT v1.3h。SAPC 和 SDAPC 的图像分析软件是 MRIcro。

2.1　任务态数据处理

任务态实验采集的是 6 个被试(4 男 2 女)的视觉刺激数据。要求被试者在 40 s 内完成 OFF-ON 两个状态的任务。在 ON 状态,要求被试者看一个蓝黄屏幕的棋盘格,在 OFF 状态,要求被试者注视屏幕的中心。整个大脑扫描层数为 36,每层厚度为 4 mm,每层体素点个数为 64 × 64,TR 为2 s,体素点大小为 3.75 mm × 3.75 mm,扫描间隔为 1 mm。数据预处理包括时间矫正和头动矫正。用 ROC 曲线[15]对结果进行优劣评估。在对 fMRI 数据分组进行 APC 时,任务态数据每个组包含100 个数据点。

2.2　静息态数据处理

静息态数据来自 Dr. James 和 J. Stewart H 共同发布的神经影像公共数据库(http://www.nitrc.org/projects/fcon_1000/),共包括 23 个被试(8 男15 女),年龄大致在 20 岁 ~40 岁之间。扫描层数为 47,每层厚度为 3 mm,每层体素点个数为 96 × 96,TR 为 2 s,体素点大小为2.67 mm × 2.67 mm,扫描间隔为3 mm。数据预处理包括时间矫正、头动矫正、标准化和平滑处理(高斯核 $FWMN$ = 8 mm)。在对 fMRI 数据分组进行 APC 时,静息态数据每个组包含 400 个数据点。

3　实验结果分析

3.1　任务态实验

在视觉刺激下,GLM,SDAPC,SAPC 得到的单被试激活区如图 1 所示,其中,$z\text{-}score$ = 2.0,p = 0.05,$FWHM$ = 0 mm。

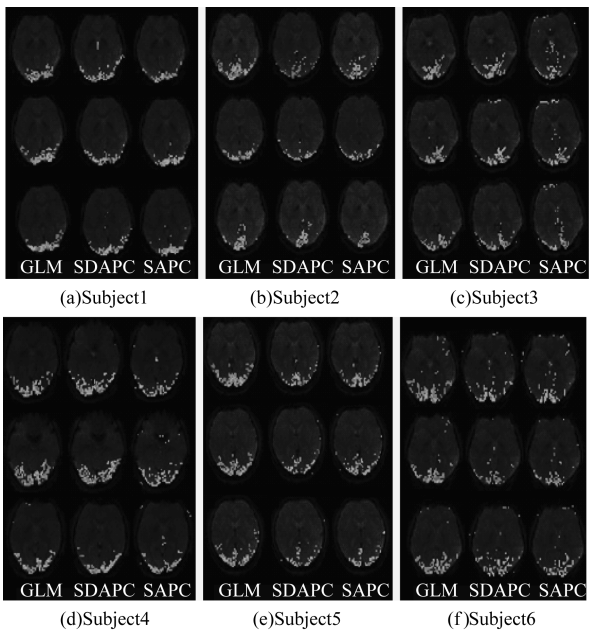

图 1　单被试的视觉区激活图比较

通过对图 1 的定性分析可知,SDAPC 和 SAPC 得到的视觉激活区与 GLM 估计得到的模板图谱大致相似。仔细观察被试 2 和被试 3 可以看出,SDAPC 检测到的激活区比 SAPC 少了很多噪声点。仔细观察被试 1 和被试 5 可以看出,SDAPC 比 SAPC 少检测到某些激活区体素点。通过以上分析得出:SDAPC 在去掉噪声点的同时也去掉了少许激活区体素,原因可能是在对相似度矩阵约减的过程中,除噪的同时也除去了某些视觉区体素点。但从整体上看,SDAPC 和 SAPC 都能检测出比较好的视觉区网络。

用 ROC 曲线对单被试激活区结果进行验证,曲线如图 2 所示。通过对图 2 的定量分析可知,SDAPC 和 SAPC 得到的脑激活区的 ROC 曲线大体相似。两条曲线都很靠近 ROC 图的左上角,假阳性率(FP)取值在 0.02 左右,真阳性率(TP)就达到 0.8。观察被试 3 和被试 4 可知,SDAPC 的 ROC 曲线比 SAPC 要好。观察被试 1、被试 2 和被试 5 可知,SAPC 的 ROC 曲线比 SDAPC 要好。通过以上分析得出:对于被试 3 和被试 4,SDAPC 得到的激活区精度比 SAPC 更高,原因可能是相似度矩阵约减起到了去噪效果。对于被试 1、被试 2 和被试 5,SDAPC 得到的激活区精度比 SAPC 要略差,原因可能是相似度矩阵约减除去了某些视觉区体素点。从整体看,SDAPC 和 SAPC 得到的视觉激活区与 GLM 模板图谱很相似,都能得出比较好的视觉区网络,验证了 SDAPC 的有效性。

单被试聚类算法运行时间如图 3 所示。通过图 3可知,SDAPC 相比 SAPC 在 6 个被试上的用时都降低很多。对于被试 1、被试 2、被试 3 和被试 5,SDAPC 运行时间比 SAPC 快 3 倍左右。对于被试 4 和 6,SADPC 比 SAPC 也要快 2 倍以上。通过以上分析得出:相比 SAPC,SDAPC 节省了算法运行时间,减少了算法的迭代次数,降低了数据的空间存储,提高了算法运行效率。

三次聚类得到的多被试激活区的 ROC 曲线如图 4 所示。多被试脑激活区图谱如图 5 所示。

通过对图 1、图 2、图 4 和图 5 分析可知,多被试脑激活区的整体噪声点比单被试少很多,且 ROC 曲线也更接近图的左上角。SDAPC 得到的脑激活区在精度上略低于 SAPC。

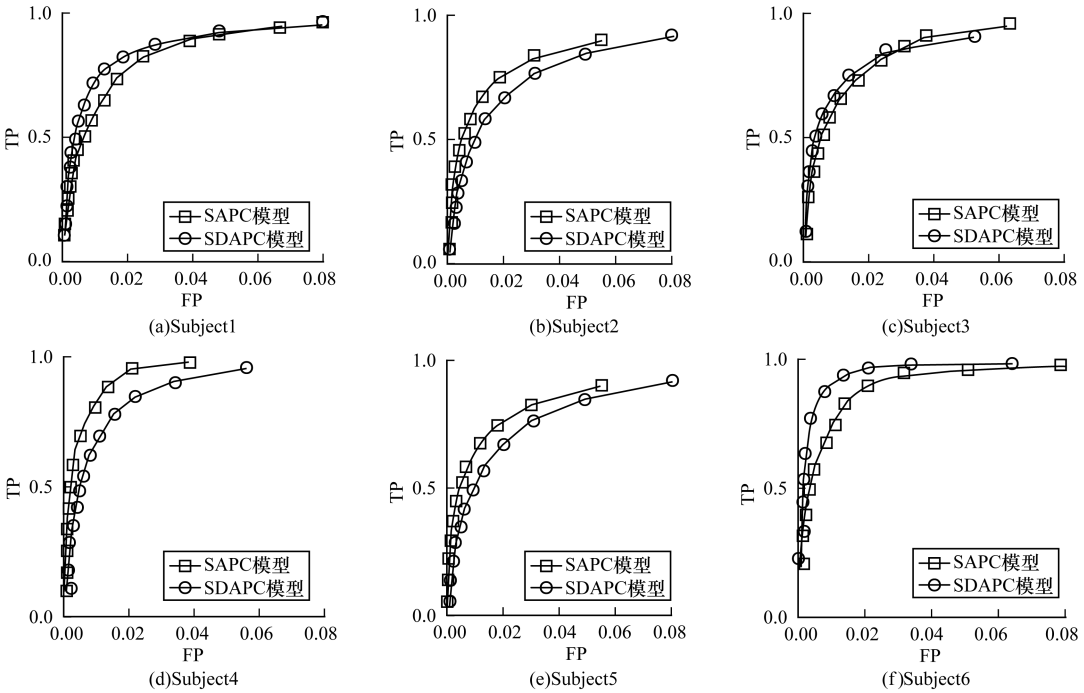

图 2 单被试脑激活区 ROC 曲线比较

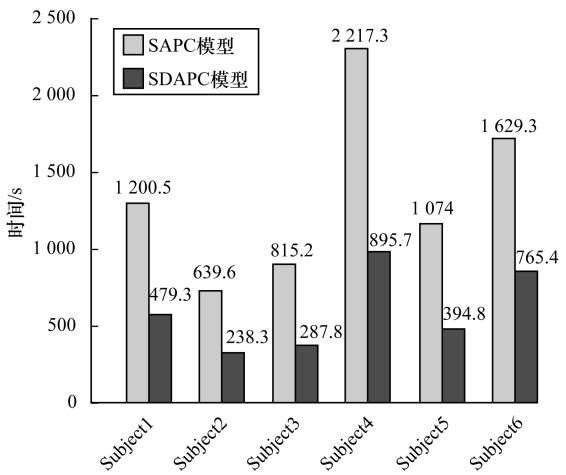

图 3 聚类时间比较

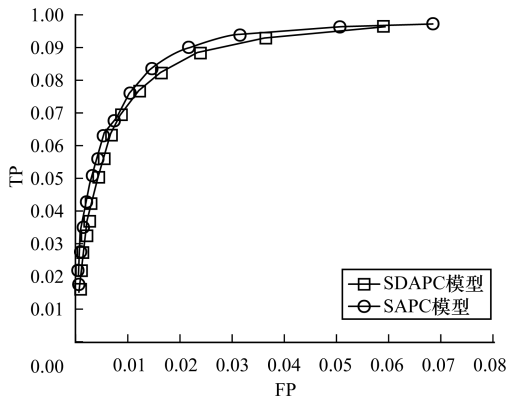

图 4 多被试激活区的 ROC 曲线比较

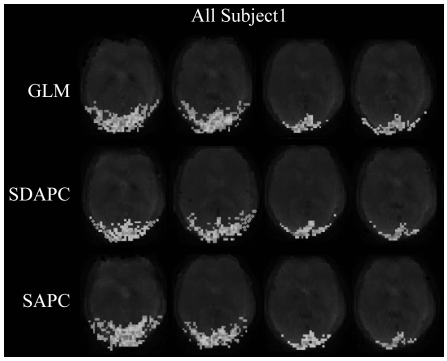

图 5 多被试的视觉区激活图比较

3.2 静息态实验

在静息态下,将 SDAPC 检测出的激活区叠加到标准脑网络模板上,得到的 9 个静息态网络如图 6 所示,其中,*FWHM* = 8 mm,*z-score* = 2.0。可以看出,SDAPC 检测出的 9 个典型的静息态脑网络[8]分别为视觉网络(VIN)、默认网络(DMN)、基底神经节区域网络(BGN)、听觉网络(AUN)、双侧感觉运动皮层网络(SMN)、左脑工作记忆网络(LWMN)、右脑工作记忆网络(RWMN)、背侧顶叶网络(DPN1)和前额叶皮层网络(DPN2)。

分析 SDAPC 检测出的 9 个静息态脑网络可知,视觉网络、默认网络、基底神经节区域网络、双侧感觉运动皮层网络、左脑工作记忆网络和右脑工作记忆网络都能比较好地检测出,且噪声点较少。SDAPC 方法能成功检测出大脑静息态下存在的 9 个脑网络,且算法运行速度相比 SAPC 也大幅提高。

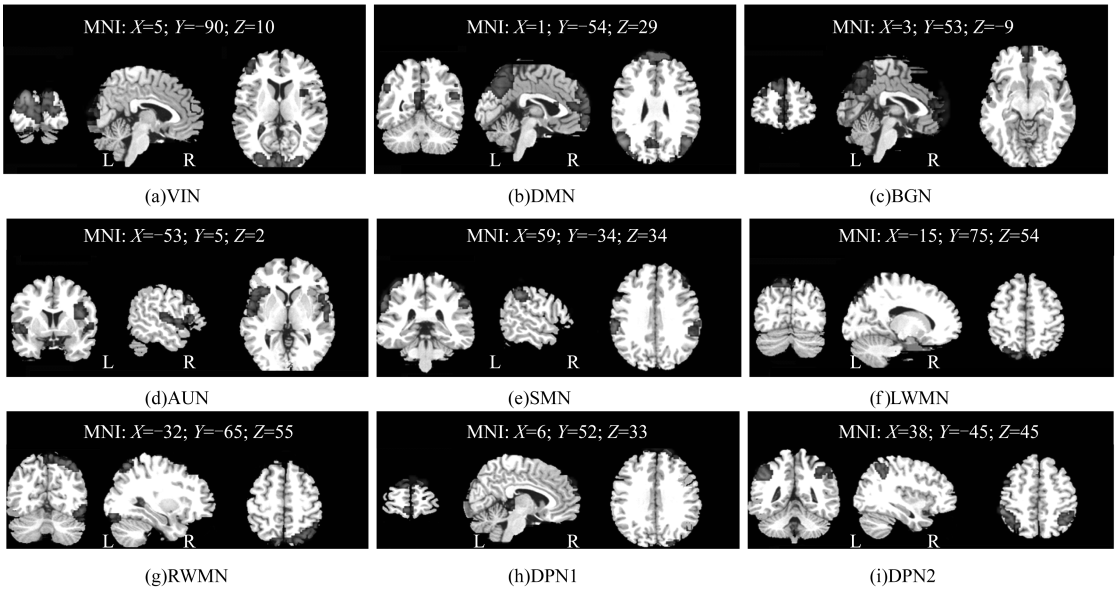

图 6　SDAPC 检测出的静息态网络

4　结束语

本文提出一种融合稀疏仿射传播聚类(SAPC)和相似度矩阵约减的仿射传播聚类算法(SDAPC)。实验结果表明,SDAPC 和 SAPC 得到的激活区的 ROC 曲线大体一致,但 SDAPC 在时间复杂度上明显优于 SAPC,这样既能节约数据的存储空间,又能简化算法复杂度。在静息态下,SDAPC 可成功检测出9 个静息态脑网络,但是 SDAPC 在激活区精度上略低于 SAPC,因此,下一步将对如何提高 SDAPC 的聚类有效性进行研究。

参考文献

[1] Frey B J, Dueck D. Clustering by Passing Messages Between Data Points[J]. Science, 2007, 315(5814): 972-976.

[2] Zhang Jiang, Tuo Xianguo, Yuan Zhen, et al. Analysis of fMRI Data Using an Integrated Principal Component Analysis and Supervised Affinity Propagation Clustering Approach[J]. IEEE Transactions on Biomedical Engineering, 2011, 58(11): 3184-3196.

[3] Zhang Jiang, Li Dahuan, Chen Huafu, et al. Analysis of Activity in fMRI Data Using Affinity Propagation Clustering[J]. Computer Methods in Biomechanics and Biomedical Engineering, 2011, 14(3): 271-281.

[4] 赵　健,唐　洁,谢　瑜. 仿射传播算法在图像聚类应用中的实现与分析[J]. 计算机应用研究, 2012, 29(10): 3980-3982.

[5] Ren Tianlong, Zeng Weiming, Wang Nizhuan, et al. A Novel Approach for fMRI Data Analysis Based on the Combination of Sparse Approximation and Affinity Propagation Clustering[J]. Magnetic Resonance Imaging, 2014, 32(6): 736-746.

[6] Wang Nizhuan, Zeng Weiming, Chen Lin. SACICA: A Sparse Approximation Coefficient-based ICA Model for Functional Magnetic Resonance Imaging Data Analysis[J]. Journal of Neuroscience Methods, 2013, 216(1): 49-61.

[7] Wang Nizhuan, Zeng Weiming, Shi Yingchao, et al. WASICA: An Effective Wavelet-shrinkage Based ICA Model for Brain fMRI Data Analysis[J]. Journal of Neuroscience Methods, 2015, 246: 75-96.

[8] 马娜娜. 基于密度的模糊聚类分析算法研究[D]. 包头:内蒙古科技大学, 2012.

[9] 王　玲,薄列峰,焦李成. 密度敏感的半监督谱聚类[J]. 软件学报, 2007, 18(10): 2412-2422.

[10] McKeown M J, Makeig S, Brown G G, et al. Analysis of fMRI Data by Blind Separation into Independent Spatial Components[J]. Human Brain Mapping, 1998, 6(3): 160-188.

[11] Kisilev P, Zibulevsky M, Zeevi Y Y. A Multiscale Framework for Blind Separation of Linearly Mixed Signals[J]. The Journal of Machine Learning Research, 2003, 4(7): 1339-1363.

[12] Donoho D L, Johnstone J M. Ideal Spatial Adaptation by Wavelet Shrinkage[J]. Biometrika, 1994, 81(3): 425-455.

[13] Donoho D L. De-noising by Soft-thresholding[J]. IEEE Transactions on Information Theory, 1995, 41(3): 613-627.

[14] Donoho D L, Johnstone I M. Adapting to Unknown Smoothness via Wavelet Shrinkage[J]. Journal of the American Statistical Association, 1995, 90(432): 1200-1224.

[15] Skudlarski P, Constable R T, Gore J C. ROC Analysis of Statistical Methods Used in Functional MRI: Individual Subjects[J]. Neuroimage, 1999, 9(3): 311-329.

编辑　金胡考