doi:10.11959/j.issn.1000-436x.2015181

改进谱聚类算法在 MCI 患者检测中的应用研究

相洁[1]，赵冬琴[2]

(1. 太原理工大学 计算机科学与技术学院，山西 太原 030024；2. 山西财经大学 实验教学中心，山西 太原 030006)

摘 要：为了利用功能核磁影像（fMRI, functional magnetic resonance imaging）数据进行轻度认知障碍（MCI, mild cognitive impairment）自动检测，对患者的 fMRI 数据进行聚类分析，得到患者大脑血氧依赖水平(BOLD, blood oxygen level dependence)的变化模式，并将异常模式用于疾病检测中。由于传统谱聚类算法需要计算相似矩阵所有的特征值和特征向量、时间与空间复杂度较高。提出一种改进的谱聚类方法，在相似矩阵的构造以及 σ 与 k 值的确定等方面进行了改进，将其用于 MCI fMRI 数据的聚类与诊断研究中。与传统谱聚类及 Nyström 算法进行的对比实验结果表明，改进的谱聚类方法可以更准确得到患者异常 BOLD 模式，分类正确率较高，且时间和空间复杂度均小于传统算法。

关键词：谱聚类；Nyström；fMRI-BOLD；轻度认知障碍；MCI 诊断

中图分类号：TP181.09 **文献标识码**：A

Improved spectral clustering algorithm and its application in MCI detection

XIANG Jie[1], ZHAO Dong-qin[2]

(1.College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan 030024, China;
2. Center of Experimental and Teaching, Shanxi University of Finance and Economics, Taiyuan 030006, China)

Abstract: In order to detect mild cognitive impairment (MCI) using functional magnetic resonance imaging (fMRI), a method based on fMRI clustering was proposed fMRI data were clustered to obtain the blood oxygen level dependence(BOLD) change model of MCI patients, then abnormal patterns were used to detect disease. The traditional spectral clustering algorithm needs to calculate all of the eigenvalue and eigenvector, so time and space complexity is higher. An improved spectral clustering method was proposed which modified the similar matrix construction method and the setting method of σ and k, and then this method was applied to clustering and detection of MCI patients. To verify the performance of the proposed method, the comparison of the clustering result, classification accuracy using traditional algorithm and Nyström is also done. The comparative experimental results show that the proposed method can get BOLD pattern more accurately, the accuracy of MCI detection is higher than the other two algorithms, and the time and space complexity are less than the traditional algorithm.

Key words: spectral clustering; Nyström; fMRI-BOLD; MCI; MCI detection

1 引言

阿尔兹海默症（AD, Alzheimer's disease）是老年期常见的慢性精神衰退性疾病，发病率较高，到 2050 年时预估全球每 85 人就有一人罹患此病[1]。AD 现有药物治疗非常有限，但早期发现、治疗能减缓疾病进程。轻度认知障碍（MCI, mild cognitive impairment)是介于正常老化和 AD 之间的过渡阶段，是 AD 的高危人群，研究表明约 44%的 MCI 患者在 3 年后转化为 AD，平均年转化率为 15%[2]，

收稿日期：2014-10-08；**修回日期**：2014-12-02

基金项目：国家自然科学基金资助项目（61170136, 61373101, 61472270, 61402318）；山西省科技攻关基金资助项目（20140321002-01）

Foundation Items: The National Natural Foundation of China (61170136, 61373101, 61472270, 61402318); The Science and Technology Iadusrid Project of Shanxi Province (20140321002-01)

而正常老年人每年仅1%~2%发展为AD。由于AD不可逆转，因此对MCI患者的临床前预警和早期干预治疗尤为重要。目前，一些医院已经利用磁共振成像（MRI, magnetic resonance imaging）进行MCI检查，但是虽然MCI患者出现了认知能力的衰退，但其脑结构并没有明显变化，因此MRI检查并不能准确进行MCI诊断。研究人员尝试利用功能磁共振（fMRI, functional magnetic resonance imaging）来研究MCI的认知衰退机制，本文将利用fMRI数据，利用聚类方法提取MCI患者特征，用于MCI诊断。

聚类是机器学习领域中的一种主要方法，可以将数据对象分组成多个簇（或类），同一簇内的数据具有很高的相似度（距离），不同簇间的数据相似度较低。传统的聚类算法（如*k*-means、FCM等）实现简单，运算效率高，应用比较成熟，但是这些算法都是建立在凸球形样本空间上，在大数据集上，尤其在非凸球型的大数据集上很难得到好的结果，经常会陷入局部最优，且最终结果受初始参数的选择影响比较大。近年来，提出的谱聚类算法可以在任意形状的样本空间上聚类，且收敛于全局最优解[3]。由于fMRI数据不一定满足凸球型样本空间，具有结构复杂、数据量大和维数多等属性，因此传统的聚类算法不适应fMRI数据的聚类，一些研究[4~8]已经将谱聚类算法用于fMRI时间序列的特征选取中。但是，传统的谱聚类算法需要根据样本数据集定义数据点之间的相似度矩阵，并计算相似矩阵的特征值和特征向量，然后选择合适的特征向量聚类不同的数据点。如果样本数量大、维度高，相似矩阵与其特征值、特征向量的计算则需要较多的存储空间。Charless Fowlkes等[9]提出Nyström谱聚类算法，根据小部分样本推断整个数据集的分组，可以降低计算的复杂性，在图片分组中研究中得到较好结果。但该算法不应定适合其他应用领域，Wen-Yen Chen[10]等将其用于分布式系统研究中发现Nyström算法聚类结果稳定性、准确性较差。

本研究拟通过一组MCI患者完成特定任务时同步采集的fMRI数据，利用谱聚类提取MCI大脑血氧变化模式，并利用异常模式进行MCI分类。由于fMRI数据包含全脑所有体素的血氧依赖水平（BOLD, blood oxygen level dependence）的时间序列，数据量大，维度高，直接使用传统谱聚类算法时计算量较大，且很难确定算法参数，聚类数量选取也较困难。本文将在传统谱聚类算法基础上，改进相似矩阵的构造以及聚类数量k与尺度参数σ的选取方法，使之适合fMRI数据的聚类。

2 谱聚类算法

2.1 经典谱聚类算法

谱聚类算法有很多不同的具体实现方法，主要有以下3个步骤。

Step1 构建样本集矩阵；

Step2 计算特征值和特征向量，构建特征向量空间；

Step3 利用*k*-means或其他经典算法对特征向量空间中的特征向量进行聚类。

定义 1 两数据样本间欧式距离的平方表示为D_{ij}，其定义如式(1)所示。

$$D_{ij}=\sum_{m=1}^{M}(x_{im}-x_{jm})^2 \quad (1\leqslant i\leqslant N,1\leqslant j\leqslant N) \tag{1}$$

定义 2 高斯核函数定义如式(2)，其中σ为尺度参数。

$$k(x,y)=\exp(-\|x-y\|^2/2\sigma^2) \tag{2}$$

经典谱聚类算法描述如下。

输入：

样本集 $S=\{s_1,s_2,s_3,s_4,\cdots,s_n\}$

聚类数量k

尺度参数σ

输出：

聚类结果

步骤：

1) 构建相似矩阵$\boldsymbol{W}\in R^{n\times n}$，矩阵中元素$W_{ij}$定义如下$W_{ij}=\exp(-\|s_i-s_j\|^2/2\sigma^2)$，其中，当$i$=$j$时，$W_{ij}$=0；

2) 构造度矩阵$\boldsymbol{D}$，矩阵主对角线上的元素$\boldsymbol{D}(i,j)$为相似性矩阵W_{ij}的第i行元素之和；然后构建构造拉普拉斯矩阵$\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$；

3) 对拉普拉斯矩阵$\boldsymbol{L}$进行特征值分解，计算特征值，找出前k个特征值所对应的特征向量$\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_k$，然后构造矩阵$\boldsymbol{X}$，$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots,\boldsymbol{x}_k]$；

4) 对矩阵$\boldsymbol{X}$的行向量进行归一化，归一化后的矩阵为$\boldsymbol{Y}$，$Y_{ij}=X_{ij}/(\sum X_{ij}^2)$；

5) 将矩阵$\boldsymbol{Y}$中的每一行看作为空间R_k中的样本，其中样本数量为n，样本维数为k，利用*k*-means

或者其他经典聚类算法对特征向量进行聚类。

6) 样本点 s_i 划分为第 j 类,当且仅当矩阵 $\boldsymbol{Y}$ 的第 i 行被划分为第 j 类。

2.2 Nyström 谱聚类算法

Nyström 是 Fowlkes 等人研究邻接矩阵和特征向量时提出的一种谱聚类算法[9]。它通过一小部分抽样的数据来逼近原始数据的特征空间，降低了计算复杂度，解决了经典谱聚类算法的内存溢出问题。

定义 3　亲和度分块矩阵定义如式(3)所示。

$$\boldsymbol{S}_d=\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{B}^{\mathrm{T}} & \boldsymbol{C}\end{bmatrix} \tag{3}$$

其中，$\boldsymbol{A}\in R^{n\times n},\boldsymbol{B}\in R^{n\times(N-n)},\boldsymbol{C}\in R^{(N-n)\times(N-n)}$。

$\boldsymbol{S}_d$ 表示一个 $N\times N$ 的相似密度矩阵，假设随机抽取 $n<<N$，$\boldsymbol{A}$ 代表样本点 $n\times n$ 的相似矩阵，$\boldsymbol{B}$ 代表 n 样本点和剩余（$N-n$）的点形成的 $n\times(N-n)$矩阵，$\boldsymbol{C}$ 代表剩余样本点$(N-n)$之间的相似矩阵。分块矩阵的目的是降低数据维数，减小计算复杂度。

Nyström 自适应谱聚类算法描述如下。

输入：

样本集 $S=\{s_1,s_2,s_3,s_4,\cdots,s_n\}$，

抽样样本数量 n，

期望的聚类分组数 k, $n>k$

输出：

聚类结果

步骤：

1) 构造矩阵 $\boldsymbol{A}\in R^{n\times n}$ 和 $\boldsymbol{B}\in R^{n\times(N-n)}$，其中$[\boldsymbol{A}\ \boldsymbol{B}]$包含 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$ 和 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 之间的相似性；

2) 计算 $a=\boldsymbol{A}\boldsymbol{1}_n,b_1=\boldsymbol{B}\boldsymbol{1}_{N-n},b_2=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{1}_n$，$\boldsymbol{D}=\mathrm{diag}\left(\begin{bmatrix}a+b_1\\ b_2+\boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{-1}b_1\end{bmatrix}\right)$，其中 $\boldsymbol{1}$ 代表全 1 的列向量；

3) 计算

$$\bar{\boldsymbol{A}}=\boldsymbol{D}_{1:N,1:N}^{-1/2}\boldsymbol{A}\boldsymbol{D}_{1:n,1:n}^{-1/2},$$

$$\bar{\boldsymbol{B}}=\boldsymbol{D}_{1:n,1:n}^{-1/2}\boldsymbol{B}\boldsymbol{D}_{n+1:N,n+1:N}^{-1/2};$$

4) 构造 $\boldsymbol{R}=\bar{\boldsymbol{A}}+\bar{\boldsymbol{A}}^{-1/2}\bar{\boldsymbol{B}}\bar{\boldsymbol{B}}^{\mathrm{T}}\bar{\boldsymbol{A}}^{-1/2}$；

5) 计算 $\boldsymbol{R}$ 的特征分解，$\boldsymbol{R}=\boldsymbol{U}_R\boldsymbol{\Lambda}_R\boldsymbol{U}_R^{\mathrm{T}}$，在 $\boldsymbol{\Lambda}_R$ 中的特征值按照递减的方式排列；

6) 构造 $\tilde{\boldsymbol{V}}=\begin{bmatrix}\bar{\boldsymbol{A}}\\ \bar{\boldsymbol{B}}^{\mathrm{T}}\end{bmatrix}\bar{\boldsymbol{A}}^{-1/2}(\boldsymbol{U}_R)_{:,1:k}(\boldsymbol{\Lambda}_R^{-1/2})_{1:k,1:k}$ 作为 $\tilde{\boldsymbol{L}}$ 的前 k 个特征向量；

7) 计算 $\tilde{\boldsymbol{V}}$ 的标准化矩阵 $\tilde{\boldsymbol{U}}$；

8) 利用 k-means 或者其他经典聚类算法聚类 $\tilde{\boldsymbol{U}}$ 的 N 行为 k；

9) 样本点 s_i 划分为第 k 类,当且仅当矩阵 $\tilde{\boldsymbol{U}}$ 的第 i 行全部被划分为第 j 类。

随机选择 n，得到聚类的正确率相当低，当 n 足够大时，会获得较稳定的结果；但是样本增多，噪音数据也随之增多，使聚类的结果有轻微的破坏。

2.3 针对 fMRI 数据特点改进的谱聚类算法

由于 fMRI 数据量大，经典的谱聚类算法不能解决内存溢出的问题，因此本文提出一种改进的谱聚类算法，用以解决内存溢出，计算复杂度高等问题。本研究中，2 个体素之间的相似度主要是衡量体素的 BOLD 变化规律是否一致，而不是体素点之间的距离，因此，本算法采用定义 4 改进相似矩阵的构造。

定义 4　相关系数

$$\rho_{X,Y}=\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}} \tag{4}$$

针对 fMRI 数据改进的谱聚类算法描述如下。

输入：

样本集 $S=\{s_1,s_2,s_3,s_4,\cdots,s_n\}$

输出：

聚类结果

步骤：

1) 针对 fMRI 数据构建矩阵 $\boldsymbol{M}$，行对应体素，列对应 BOLD-fMRI 变化率，计算所有体素之间的相关系数得到矩阵 $\boldsymbol{R}$；

2) 变换 $\boldsymbol{R}$ 为稀疏矩阵 $\boldsymbol{C}$,每 4 个相邻体素保留 1 个作为样本；

3) 根据稀疏矩阵构造亲和矩阵 $\boldsymbol{W}$，亲和矩阵被定义为 $W_{ij}=\begin{cases}\exp\left(-\dfrac{C^2(i,j)}{2\sigma_i\sigma_j}\right), & i\neq j\\ 0, & i=j\end{cases}$

其中，$C(i,j)$是 2 个体素向量的相关系数；

4) 定义 $\boldsymbol{D}$ 为对角矩阵，$D_{ij}=\sum_{j=1}^{n}W_{ij}$,构造拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$；

5) 计算 $\boldsymbol{L}$ 特征值，对其按照降序的排列方法依次为 $\lambda_1\geqslant\lambda_2\geqslant\lambda_3\geqslant\cdots\lambda_n$，计算特征值之间的差值 $g_i=\lambda_i-\lambda_{i+1}$，依次排列为 $G=\{g_1,g_2,g_3,\cdots,g_{n-1}\}$，

k=argmax{G}；

6) 得到 k 个特征向量依次为 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k$，构成矩阵 $\boldsymbol{X}=|\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_k|$；

7) 对 $\boldsymbol{X}$ 的行进行单位化形成矩阵 $\boldsymbol{Y}$，$Y_{ij}=X_{ij}/(\sum X_{ij}^2)^{1/2}$；

8) 处理 $\boldsymbol{Y}$ 的每行作为 R_k（k 维子空间）中的点，通过 k-means 或其他聚类算法对它们进行聚类。

改进的谱聚类算法主要对相似矩阵的构造、σ 的选取、k 值的确定做了改进。其中相似矩阵的改进体现在步骤 1)中，相似矩阵需要真实反映数据点之间的近似关系，保证相近点间的相似度更高，相异点之间的相似度更低。传统谱聚类算法中相似程度使用欧式距离，改进的算法使用定义 4 所示的相关系数。

为了解决计算速度问题，步骤 2)中根据 fMRI 数据特点，将相似矩阵变换为稀疏矩阵。计算过程是将所有数据点的相关系数进行排序，保留重要的相关系数，剔除比指定阈值范围更小的数据点之间的关系，这个过程降低了数据维数，保留了重要数据信息，解决内存溢出，提高了运算效率。研究中根据认知神经解剖特点，每 4 个位置相邻的体素，只保留一个。

σ 值的改进体现在步骤 3)中，传统的谱聚类算法中需要凭经验预先设定不同的 σ 值，分别进行聚类，将聚类结果中最好的 σ 作为参数，这种方法计算繁琐、耗费时间长。改进的谱聚类算法中采用自适应的方式进行选择，将 $A_{ij}=\begin{cases}\exp\left(-\dfrac{C^2(i,j)}{2\sigma_i\sigma_j}\right), & i\neq j\\ 0, & i=j\end{cases}$ 中的 σ_i 定义为 2 个数据点间相关系数的均值。

聚类数量 k 的确定主要体现在步骤 5)中，经典 NJW 谱聚类中 k 值的选取是手工选择，结合经验通过多次计算确定合适的值，会消耗大量的时间。目前，一些研究人员利用特征距或特征差值自动确定聚类数量[11]，但是此类研究还较少。本文利用特征距和传统手工选择相结合的方法确定 k 值，根据矩阵的摄动原理，通过特征距的方法自动确定聚类数目。

3 方法与数据

3.1 实验数据

数据采自北京宣武医院，研究邀请 21 位被试参与实验（11 位 MCI 患者，10 位正常被试），正常组和 MCI 组的年龄、性别均没有显著差异。任务采用事件相关设计[12]，实验要求被试完成一些 4×4 Sudoku 游戏任务，刺激任务与刺激呈现方式如图 1 所示。2 s “*”提示之后，出现 4×4 Sudoku 任务。要求被试在任务呈现阶段获得答案后尽可能快地按键；当被试按键或呈现时间超过 20 s 时任务消失，出现 2 s 空白棋盘，要求被试在此这段时间口述报告“？”的正确答案，之后让被试休息 10 s。每个被试均需完成 60 个任务（30 个简单和 30 个复杂任务）。被试完成任务的同时使用 3.0T MR 扫描仪（Siemens Trio+tim，Genmeny）核磁设备采集 fMRI 数据，每 2 s 采集一次被试的全脑 BOLD 数据。

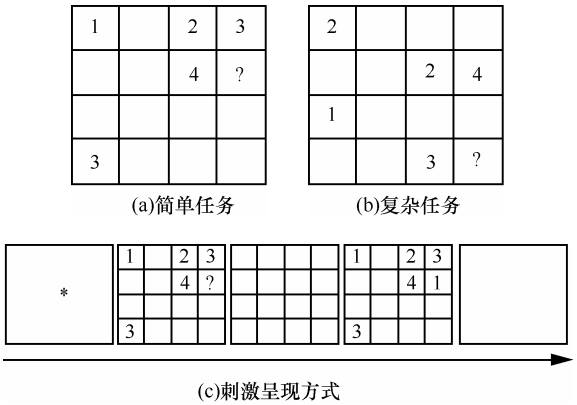


图 1 刺激任务与刺激呈现方式

每个被试采集到的 fMRI 时间为 34 min，因此每个被试均得到 1 020 个 fMRI 文件，每个 fMRI 文件都包含了在某个时刻被试全脑所有体素的 BOLD 情况，每个体素只有 9 mm^3，全脑大约包含 5 万个体素。可以看出，本项研究使用的 fMRI 数据量大，维度高。

3.2 fMRI 数据预处理

预处理过程采用 DPARSF (data processing assistant for resting-state fMRI)[13]软件对数据进行预处理，包括头动校正、时间片配准、与标准脑配准、高斯平滑（半高全宽为 6 mm）去噪。

被试在执行任务时的 BOLD 变化情况可以反应被试的认知变化模式，本研究将每个任务的休息和 2 s 注视时间的 BOLD 水平作为基线，计算其他时间 BOLD 水平相对于基线水平的变化情况。图 2 为某个任务执行过程中的 BOLD 变化率曲线。其中横坐标代表时间，纵坐标代表 BOLD 变化率，第 1、2 时间点的 BOLD 水平为任务基线，3~10 时间点为相对于基线的 BOLD 变化率。

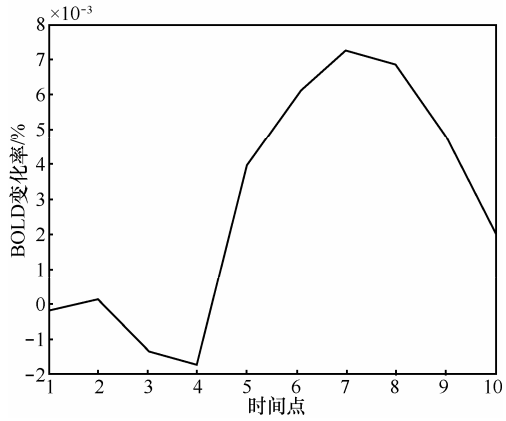


图 2　BOLD 变化率曲线

3.3　BOLD 模式聚类

大脑的不同体素属于不同脑功能区，承担着不同认知功能。由于 MCI 患者的认知能力衰退，某些体素的 BOLD 变化规律和正常人有所不同。

本研究首先通过聚类算法找到 MCI 患者和正常被试在完成 4×4 Sudoku 任务时所出现的所有 BOLD 模式，并通过 BOLD 变化峰值及峰值出现时间等因素判断 MCI 患者的异常 BOLD 模式。这些体素上出现的异常 BOLD 模式即是 MCI 的异常特征，可以用来进行 MCI 患者的诊断。

为了对比研究，本文分别采用改进的谱聚类算法，传统谱聚类算法、Nyström 算法进行 BOLD 模式聚类。

3.4　MCI 与健康被试分类

呈现异常 BOLD 模式的体素即为 MCI 患者出现认知受损的脑区，这些脑区可以作为感兴趣区域（ROI, region of interest）。计算 MCI 患者和正常被试每个 ROI 包含体素的 BOLD 变化率曲线，提取曲线的斜率，最大值、最小值、最大最小差值以及 BOLD 累计变化等作为特征，利用 SVM 分类算法进行 MCI 分类器训练。研究中，使用 lib-svm[14]工具包进行分类器的训练和测试，采用 10 折交叉验证的方法来验证分类器的性能。

4　实验结果与分析

4.1　BOLD 模式聚类结果

由于数据量大，出现内存溢出，传统算法无法得到聚类结果。Nyström 算法进行 BOLD 模式聚类时，没有出现内存溢出。图 3 是 Nyström 谱聚类算法得到的 BOLD 模式，图 4 是改进的谱聚类算法得到 BOLD 模式。

可以看到患者与正常人在完成简单或复杂任务时的 BOLD 变化模式不尽相同，这些改变可以通过 BOLD 峰值及其出现的时间得到体现。

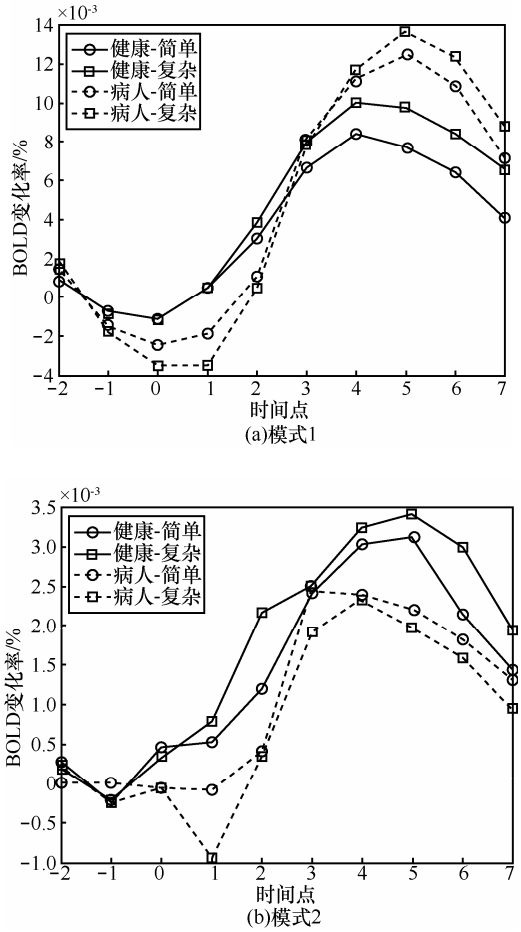
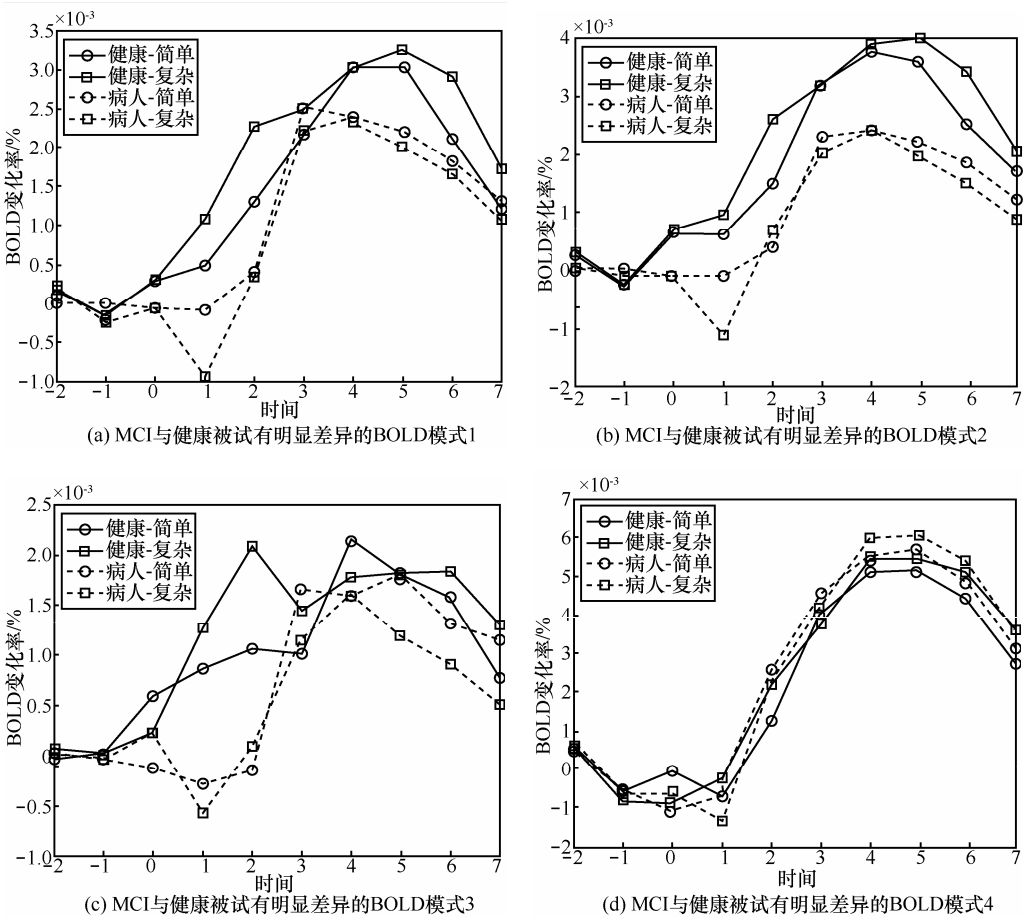


图 3　Nyström 算法得到 2 种有意义的 BOLD 模式

在图 4(a)中病人和健康人的 BOLD 变化率有明显的差异，病人的 BOLD 变化时间略早于正常人，而且病人的峰值到达时间明显早于正常人。图 4(b)中健康人和病人的 BOLD 变化率值有明显的区分，而且病人的峰值到达时间也略早于正常人；图 4(c)中病人为负激活，健康人为正激活，可以明显地区分健康人和病人；图 4(d)没有明显差异。

4.2　异常 BOLD 模式对应脑区

根据聚类结果得到的异常 BOLD 变化模式对应的体素坐标，将其映射到标准脑图谱中。图 5 是 Nyström 得到异常 BOLD 模式对应的 ROI，图 6 是改进谱聚类算法对应的 ROI。



(a) MCI与健康被试有明显差异的BOLD模式1

(b) MCI与健康被试有明显差异的BOLD模式2

(c) MCI与健康被试有明显差异的BOLD模式3

(d) MCI与健康被试有明显差异的BOLD模式4

图 4 改进算法得到 BOLD 模式

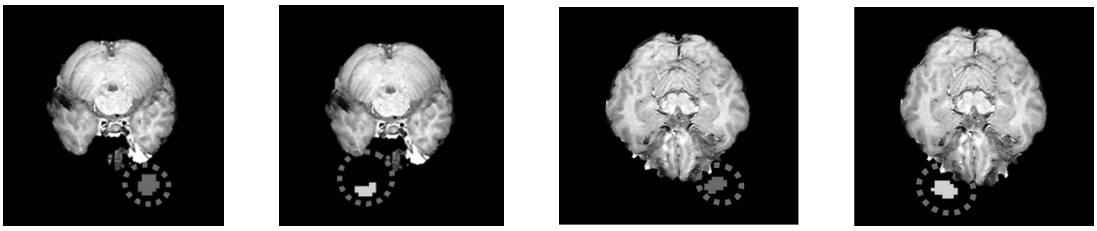


图 5 Nyström 算法得到的 ROI

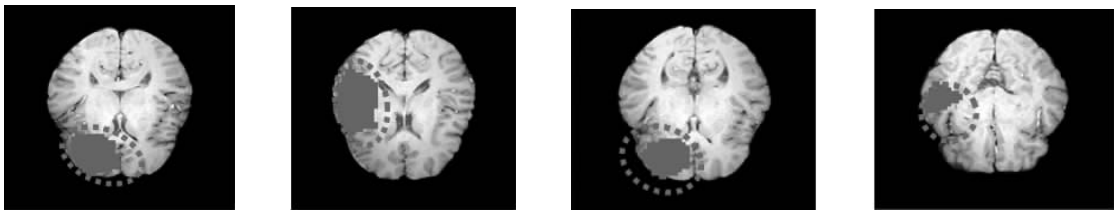


图 6 改进算法得到的 ROI

图 5 结果表明，Nyström 算法得到的异常 BOLD 模式所对应的 ROI 部分位于大脑皮层之外，显然不符合认知规律。

图 6 结果表明，改进的谱聚类算法中异常 BOLD 模式对应脑区主要包括前额叶和颞叶等脑区，这些区域主要负责推理、记忆等。认知研究表明颞叶和前额是 MCI 患者认知受损的主要脑区，本结果与这些研究一致。

4.3 分类正确率

提取 ROI 的 fMRI 时间序列，计算斜率，最大值、最小值、最大最小差值以及 BOLD 累计变化，形成特征向量，将特征向量连同被试类型（患者或

健康）一起用于 SVM 分类模型的训练。表 1 为 2 种聚类算法对应的分类正确率。其中 Nyström 算法对应的正确率为 71%，改进算法对应的正确率为 80%，很明显，改进的谱聚类算法得到的分类正确率较高。

与已有使用相同数据集的分类研究相比，文献[15]筛选特征采用了单体素之间的特征对比方法，筛选速度较慢。本研究采用聚类方法，较快较准确地找到 MCI 患者的异常 BOLD 模式来作为分类特征，明显提高了分类正确率。

表 1　Nyström 与改进谱聚类算法的分类正确率

方法	正确率
Nyström	71%
改进算法	80%

4.4　算法比较

表 2 给出了改进的谱聚类算法、Nyström 谱聚类算法和经典 NJW 谱聚类算法在时间和空间复杂度的对比，表 3 给出了 3 种聚类算法的运行时间对比。

表 2　　3 种聚类算法的复杂度对比

算法	时间	空间
经典谱聚类算法	计算 $N\times N$ 维的相似矩阵 $\boldsymbol{W}$，计算 $n\times n$ 的对角矩阵 $\boldsymbol{D}$，所有特征值和特征向量	需要存放 $n\times n$ 维的数组和所有的特征向量
Nyström 算法	只需要计算矩阵维数为 $n\times n$ 的 $\boldsymbol{A}$ 和 $n\times(N-n)$的 $\boldsymbol{B}$，高斯核函数具有正定性，直接估计 $\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$ 的主特征向量	不需要将相似矩阵 $\boldsymbol{W}$ 放到内存中，估计是一个理论过程，只需存放 $\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$ 的主特征向量
改进算法	计算相似矩阵，对所有特征值进行排序。计算特征距最大的特征值对应的特征向量	存放稀疏矩阵

相对于传统算法与 Nyström 算法，改进算法降低了时间与空间复杂度，在 fMRI 时间序列这种大的数据集上运行，没有出现内存溢出，而且运行时间（17.2 min）略少于Nyström谱聚类算法（25.23 min）。这个结果表明改进的谱聚类算法解决了传统谱聚类算法在大数据处理过程中出现的内存不足问题，而且弥补了 Nyström 谱聚类算法在选取样本数据过程中存在的片面性。

5　结束语

谱聚类算法作为一种新的分析算法，克服了传统聚类算法的缺点，可以在任意的形状上聚类，便于 fMRI 数据聚类分析，但是传统算法计算复杂度大，容易出现内存溢出以及聚类参数（聚类数量 k 及 σ）选取复杂的问题。本文针对存在的这些问题，在经典谱聚类算法的基础上做了 3 方面改进，主要是相似矩阵的构造、k 值的选取和 σ 的确定。改进的谱聚类算法不仅解决了经典谱聚类算法在大数据集上会出现内存溢出的问题，也避免了 Nyström 谱聚类算法正确率低的问题，提高了运算效率，节省了内存空间。该方法应用于 MCI 数据的 BOLD-fMRI 模式聚类后，找到了异常 BOLD 模式所对应的脑区，这些脑区均为 MCI 关键脑区，使用这些异常模式进行 MCI 分类的正确率为 80%，显著高于随机水平，这种 MCI 诊断方法可以为MCI的功能核磁影像检查提供一定的参考。

参考文献：

[1] BROOKMEYER R, JOHNSON E, ZIEGLER G K, *et al*. Forecasting the global burden of Alzheimer's disease[J]. Alzheimers Dement, 2007, 3 (3): 186-191.

[2] MISRA C, FAN Y, DAVATZIKOS C. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI[J]. NeuroImage, 2009, 44:1414-1422.

[3] 蔡晓妍，戴冠中，杨黎斌. 谱聚类算法综述[J]. 计算机科学，2008,35(7):14-18.
CAI X Y, DAI G Z, YANG L B. Survey on spectral clustering algorithms[J]. Computer Science,2008,35(7):14-18.

[4] JORDAN M I, NG A Y, WEISS Y. On spectral clustering: analysis and an algorithm[A]. Proceedings of the 14th Advances in Neural Information Processing Systems (NIPS 2002)[C]. Cambridge, MA, 2002. 849-856.

[5] YEO B T, OU W. Clustering fMRI time series[EB/OL].http://people.csail.mit.edu/ythomas/uopublished/6867fMRI.pdf, 2004.

[6] WU C. Feature selection for fMRI classification[J]. Program of Computational Biology Carnegie Mellon University Pittsburgh, PA 15213.

[7] WANG C, TIAN J, CHEN S, *et al*. Image segmentation using spectral clustering[A]. 2012 IEEE 24th International Conference on Tools with Artificial Intelligence. IEEE Computer Society[C]. 2005. 677-678.

[8] ALKAN S, YARMAN V F T. Localization of semantic category classification in fMRI images[A]. Signal Processing and Communications Applications Conference (SIU)[C]. 2014. 2178-2181.

[9] FOWLKES C, BELONGIE S, CHUNG F, *et al*. Spectral grouping

using the Nystrom method [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(2): 214-225.

[10] CHEN W Y, SONG Y, BAI H, *et al*. Parallel spectral clustering in distributed systems [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(3): 568-586.

[11] 孔万增, 孙志海, 杨灿等. 基于本征间隙与正交特征向量的自动谱聚类[J]. 电子学报, 2010, 38(8):1880-1891.
KONG W Z, SUN Z M, YANG C. *et al*. Automatic spectral clustering based on eigengap and orthogonal eigenvector[J]. Acta Electronica Sinica, 2010, 38(8):1880-1891.

[12] 相洁.启发式问题解决认知神经机制及 fMRI 数据分析方法研究[D]. 太原理工大学, 2010.
XIANG J. Study of Cognitive Neuroscience Mechanism of Heuristic Problems Solving and Methods of fMRI Data Analysis[D]. Taiyuan University of Technology,2010.

[13] YAN C G, ZANG Y F. DPARSF: a MATLAB toolbox for “pipeline” data analysis of resting-state fMRI [J]. Frontiers in Systems Neuroscience, 2010, 5(4):13.

[14] CHANG C C, LIN C J. LIBSVM: a library for support vectormachines[EB/OL].http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[15] 吕艳阳, 相洁. 基于 SVM 的 fMRI 数据分类及 MCI 诊断应用[J]. 计算机工程与设计, 2013, 34(9): 3313-3317.
LV Y Y, XIANG J. fMRI data classification based on SVM and its application in diagnosis of MCI[J]. Computer Engineering and Design, 2013, 34(9): 3313-3317.

作者简介：

相洁（1970-），女，山西太原人，太原理工大学副教授，主要研究方向为数据挖掘、脑信息学、人工智能及其应用。

赵冬琴（1984-），女，山西阳泉人，山西财经大学硕士生，主要研究方向为智能信息处理。