文章编号：0253-374X(2016)09-1450-08 DOI:10.11908/j.issn.0253-374x.2016.09.021

基于粒子群-支持向量机的时间序列分类诊断模型

张 涛[1,3]，张明辉[1]，李清伟[2]，张玥杰[4]

(1. 上海财经大学 信息管理与工程学院，上海 200433；2. 同济大学附属同济医院，上海 200065；
3. 上海市金融信息技术研究重点实验室(上海财经大学)，上海 200433；
4. 复旦大学 计算机科学技术学院，上海市智能信息处理重点实验室，上海 200433)

摘要：构建一种基于粒子群算法-支持向量机(PSO-SVM)的磁共振功能成像(fMRI)时间序列分类诊断模型，通过针对脑区多维时间序列数据的深层次分析实现病症患者和健康者的准确判断与区分，为面向fMRI时间序列数据的病症诊断和预测提供有效科学依据. 该方法在以下4个方面不同于其他已有相关研究工作：(1) 构建基于自回归模型的脑区多维时间序列数据特征表示；(2) 构建基于支持向量机模型的脑区多维时间序列数据分类机制；(3) 构建基于粒子群算法的分类学习参数寻优策略；(4) 建立融合上述特征表示、优化分类与参数优选模式的fMRI时间序列数据分类诊断模型. 通过以精神抑郁症作为实证分析的具体案例，所提出分类诊断模型已取得良好实验效果，展示出其有效性与合理性.

关键词：fMRI多维时间序列；分类诊断；自回归模型；支持向量机(SVM)；粒子群算法(PSO)

中图分类号：TP391 **文献标志码**：A

Time Series Classification Diagnosis Model based on Partical Swarm Optimization and Support Vector Machine

ZHANG Tao[1,3], *ZHANG Minghui*[1], *LI Qingwei*[2], *ZHANG Yuejie*[4]

(1. School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China; 2. Department of Psychiatry, Tongji Hospital of Tongji University, Shanghai 200065, China; 3. Shanghai Key Laboratory of Financial Information Technology (Shanghai University of Finance and Economics), Shanghai 200433, China; 4. School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433, China)

Abstract: This paper presents a fMRI time-series classification diagnosis model based on particle swarm optimization-support vector machine (PSO-SVM), which achieves more accurate judgments and distinctions between patients and healthy individuals by deeply analyzing multi-dimensional time-series data of brain regions. This approach is significantly different from the other existing related research work in four aspects as follows: Constructing the feature representation for multi-dimensional time-series data of brain regions based on the auto-regressive (AR) model; Constructing the classification scheme for multi-dimensional time-series of brain regions based on the support vector machine (SVM) model; Constructing the parameter optimization strategy for the classification learning based on the particle swarm optimization (PSO) algorithm; Constructing the classification diagnosis framework for fMRI time-series data by integrating the above feature representation, optimized classification and parameter optimization patterns. With the mental depression disorder (MDD) as a specific case of empirical analysis, our classification diagnosis model has obtained very positive numerical computation results.

Key words: fMRI multi-dimensional time series; classification diagnosis; auto-regressive (AR) model; support vector machine (SVM); particle swarm optimization (PSO)

大量的诊断记录、医学影像、化验与测量结果、生物基因序列等信息组成十分庞大的医疗诊断时间序列数据库[1]，有关医学诊断时间序列数据分析与挖掘的应用越来越广泛. 特别是，近年来涉及认知神经科学的医疗诊断通常采用各种脑成像技术所生成的典型时间序列数据来进行和大脑神经活动相关的

收稿日期：2015-07-12
基金项目：国家自然科学基金(71171126，61572140)；教育部高等学校博士学科点专项科研基金(20130078110001)；上海市科学技术委员会科技创新行动计划资助项目(16511104704)；同济大学青年优秀人才培养计划(1508-219-040).
第一作者：张 涛(1970—)，男，工学博士，教授，博士生导师，主要研究方向为数据挖掘、物流建模与优化、智能优化算法. E-mail：taozhang@mail.shufe.edu.cn
通讯作者：李清伟(1977—)，男，医学博士，副主任医师，主要研究方向为神经影像学、智能诊断. E-mail：lianocd@tongji.edu.cn

诊断决策研究，如脑电、脑磁、正电子断层扫描、磁共振成像及磁共振功能成像（functional magnetic resonance imaging，fMRI）[2]. 有关 fMRI 时间序列数据的分析与挖掘大都依赖于时间序列数据的相似性或不相似性的度量，但简单的等式或不等式策略在时间序列数据上作用微乎其微且时间序列往往很长，这就使得针对时间序列数据的处理问题更容易陷入“维度灾难（dimensional curse）”[3]. 实现面向 fMRI 时间序列数据信息的深层次分析与挖掘进行有效的分类诊断，已成为医疗数据研究与发展的新兴热点.

fMRI 时间序列数据能充分展示大脑不同活动的变化，不仅包含大脑活动的时间信息，同时也包含大脑内部的空间特征，亟需高效成熟的分析与挖掘方法去捕获其在时间和空间上的特点. 然而，目前基于 fMRI 时间序列数据挖掘的分类研究大部分只关注其中单一种类的时间信息或者空间特征，鲜有针对庞大 fMRI 时间序列数据集中特有时空模式特点的科学定量研究[4]. 为更加有效地实现 fMRI 时间序列数据有效分类诊断，亟需解决三个关键问题：① fMRI 时间序列数据的合理特征表示，以特征化所蕴含的丰富时空属性信息；② 融合 fMRI 时间序列数据特征的优化分类诊断机制，以标识时间序列数据与具体疾病类别之间的关联；③ 针对 fMRI 时间序列数据分类诊断的参数调优策略，以保证分类机制的准确性与时间效率. 为解决第一个问题，非常重要于利用大规模 fMRI 时间序列数据进行深度分析与挖掘，以全面获取有益特征信息. 为解决第二个问题，非常关键于探索一种合适 fMRI 时间序列数据的分类诊断模型，以充分利用合理特征表述充分挖掘 fMRI 时间序列数据属性与疾病类别之间的内在相关性. 为解决第三个问题，非常有必要构建新型的参数选择模式，以实现有效支撑分类诊断性能和效率的最优参数设置.

本文提出一种基于粒子群算法-支持向量机（particle swarm optimization-support vector machine，PSO-SVM）的 fMRI 时间序列分类诊断模型，通过针对脑区多维时间序列数据的深层次分析实现病症患者和健康者的准确判断与区分，该方法在以下方面显著不同于其他已有相关研究工作：① 构建基于自回归（auto-regressive，AR）模型的脑区多维时间序列数据特征表示，合理表征这一特殊时间序列数据的内在属性信息；② 构建基于支持向量机（support vector machine，SVM）模型的脑区多维时间序列数据分类机制，实现 fMRI 时间序列数据有效模式与关系的深度分析与挖掘；③ 构建基于粒子群算法（particle swarm optimization，PSO）的分类学习参数寻优策略，获取最佳参数组合以优化分类模型；④ 建立融合上述模式的 fMRI 时间序列数据分类诊断模型，为病症的科学诊断和预测形成定量指导和参考指标. 通过以精神抑郁症作为实证分析的具体案例，所提出分类诊断模型已取得良好的实验效果，展示出其有效性与合理性.

1　相关工作

多维时间序列不同于通常的一维时间序列或者单变量时间序列，在具体应用场景中一个时刻对应多个实数变量或者描述属性而非简单的一维实值变量，同时其相互之间的关系不再局限于变量与时间之间的相关性，变量与变量之间也会具有一定的联系[5-6]. 近些年来，面向医学领域的脑电信号序列、磁共振成像 fMRI 时间序列等特有的脑区多维时间序列数据，越发体现出其在病症医学诊断和预测中非常重要的参考价值和借鉴意义. 然而，脑区多维时间序列除具有数据规模较大、随时间不断变化、较高维度、结构较为复杂等特点外，还具有医学信息所独有的特殊性与复杂性，因此，面对大量脑区多维时间数据信息，探究合理有效的分类诊断和预测模型作为医学诊疗的辅助工具十分必要.

目前，融合基于假设驱动和基于数据驱动分类方法优势的分类模式，已成为针对 fMRI 多维序列数据分类诊断研究的热点. 基于假设驱动的分类方法，以具体感兴趣区域为基础设定相关脑区，通过分析特定脑区的时间序列信息来构架分类模型. 基于数据驱动的分类方法，以经过数据采集获取的数据样本为基础提取相关特征属性信息，通过基于特征属性描述的分析与挖掘来构架分类模型. 两者相辅相成，优势互补，以感兴趣区域为基础针对病症关联的多脑区多维时间序列信息进行特征提取，筛选出与病症相关的关键属性关系，相对较为全面地反映 fMRI 多脑区多维时间序列信息的内在有价值数据特征，降低个体差异带来的误差影响. Norman 等利用 SVM 训练分类器，可通过 fMRI 数据判断被测试者正在观看的是鞋子还是瓶子，也可判断被测试者观察的图片是人脸还是物体[7]. Costafreda 等和 Fu 等分别将 SVM 分类器应用到结构像和功能像 fMRI 数据中，正常对照组和抑郁症组的分类准确率分别

为67%和86%[8]. Chupin等提出一种基于概率和解剖经验的全自动海马分割方法,针对老年痴呆(Alzheimer's disease, AD)患者、轻度认知障碍(mild cognitive impairment, MCI)患者和正常对照组的fMRI图像进行海马分割而得到海马容积,然后利用SVM算法进行两组间分类研究[9]. Gong等以灰质及白质组间差异为特征,使用SVM算法来区分难治性(refractory depressive disorder)和易治性(non-refractory depressive disorder)抑郁症,其准确率分别达到65.22%和76.09%[10]. 相洁等使用全脑体素实现分类,以预测人脑所在执行的高级思维状态[11]. Mitchell等对比研究不同的分类算法,包括支持向量机(SVM)、动态贝叶斯网络(dynamic bayesian networks, DBNs)、高斯朴素贝叶斯(Gaussian Naive Bayes, GNB)和k2最近邻分类算法(k2 nearest neighbor, kNN)等,体现出SVM在解决小样本、非线性及高维模式识别方面具有一定优势[12-14]. 一些相关研究针对脑疾病的多种MRI数据进行多元模式分类开展,包括结构像(如Ti加权像)、功能磁共振成像fMRI、静息态功能磁共振成像、及扩散张量成像等[15-16]. 另外,还有一些相关研究在脑疾病分类研究中结合MRI中多种模态的图像数据,对多种精神疾病进行多元模式分类分析,如阿尔茨海默病、精神分裂症、抑郁症及抑郁狂躁型忧郁症(bipolar disorder)等[17-18].

由以上分析,针对fMRI脑区多维时间序列数据分类诊断研究,应更充分考虑以下方面:① 针对fMRI时间序列数据的深层次分析以挖掘有效特征表示——应通过深度分析与挖掘fMRI时间序列数据,探索蕴含有价值属性信息的特征空间构建,为分类诊断构架有益特征描述基础;② 针对fMRI时间序列数据与病症间相互关联性的合理度量——从fMRI时间序列数据特定属性信息的探究、利用和融合的角度,十分有必要组合其特有特征表述与适宜的分类决策模型以构建适用于脑区多维时间序列数据和病症关联度评测的分类诊断联合模型;③ 针对fMRI时间序列数据分类诊断的参数设置优化——构架分类诊断的参数优化设置模式,能够在很大程度上控制整个分类决策过程的时空代价在一个合理范围之内,有效促进分类诊断性能和效率的进一步改进与提升.

2 基于AR模型的fMRI时间序列数据特征表示

fMRI脑区多维时间序列数据中通常存在部分不相关或冗余的维度,针对多维时间序列分析必须考虑各维度间的联系,特别是不仅将时间序列看作是空间向量而关注维度之间的数量关系,同时也要充分利用隐藏在多维时间序列中的时间信息. 若直接对原始时间序列进行特征提取则会大大降低分类诊断性能,有必要寻求一种有效特征表示模式来剔除不相关维度.

大量研究表明,AR模型的自回归系数对状态变化规律敏感,采用AR模型的自回归系数作为特征向量来分析时间序列邻域相关的变化十分有效. 但由于AR模型只能针对平稳时间序列建模,因此在建模之前必须对时间序列进行平稳性检验,对于模型阶数的确定则只能满足一定置信范围. 为此,特别考虑使用AR模型来拟合每个脑区的时间序列,其中使用平滑先验方法(smoothness prior approach, SPA)和经验模式分解(empirical mode decomposition, EMD)方法对非平稳的原始时间序列去除趋势项,对残差项进行独立同分布(independent identical distribution, IID)平稳性检验,检验通过后根据其自相关系数和偏自相关系数进行定阶,并利用Burg算法进行AR模型参数的估计而最终得到自回归方程.

自回归模型是拟合平稳时间序列的模型,由此针对每个样本脑区多维时间序列$\{Z_i\}_{i=1}^{N}$,对每一个脑区变量的时间序列$Z_i=(z_1, z_2, \cdots, z_t)$建立如下$q$阶自回归模型$A_R(q)$:

$$z_t=\varphi_0+\varphi_1 z_{t-1}+\varphi_2 z_{t-2}+\cdots+\varphi_p z_{t-q}+\varepsilon_t \quad (1)$$

其中,包含三个限制条件:(1) 模型的最高阶数为q,即$\varphi_q\neq 0$;(2) 随机干扰序列ε_t为零均值的白噪声序列,即$\varepsilon_t\sim W_N(0, \sigma_\varepsilon^2)$;(3) 当期的随机干扰与过去的序列值无关,即$Ez_s\varepsilon_t=0, s<t$. z_t的概率分布公式为

$$p(z\mid\sigma^2, z_{t-1}, \cdots, z_{t-q})=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(z_t-\varphi_1 z_{t-1}-\varphi_2 z_{t-2}-\cdots-\varphi_q z_{t-q})}{2\sigma^2} \quad (2)$$

则q阶回归模型$A_R(q)$产生拟合时间序列x_t的似然概率为

$$L(z\mid\sigma^2, \varphi_1, \cdots, \varphi_q; z_{t-1}, \cdots, z_{t-q})=\log p(z_1, \cdots, z_q)+\sum_{t=q+1}^{T}\log p(z\mid\sigma^2, z_{t-1}, \cdots, z_{t-q})=\log p(z_1, \cdots, z_q)+\frac{1}{2}(T-q)\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{t=q+1}^{T}(z_t-\varphi_1 z_{t-1}-\cdots-\varphi_q z_{t-q})^2 \quad (3)$$

需要注意的是,fMRI各脑区时间序列数据的

AR 阶数并不相等，而后续分类诊断模型要求每个时间序列样本均提供相同数目的特征属性值. 为避免丢失过多的邻域信息且有效表征原始脑区时间序列的特征属性，不能直接利用自回归参数作为时间序列样本特征信息，而上述所构建的 AR 模型拟合概率模式成为实现各脑区时间序列相同阶数特征属性表示的适宜度量.

3　基于 SVM 模型的 fMRI 时间序列数据分类机制

fMRI 脑区多维时间序列分类问题可描述为：给定一个 fMRI 多维时间序列数据集$\{T_i \mid i=1,\cdots,M+N\}$，其中每个时间序列样本 T_i 属于 $\mathbf{R}^{m*n}$，训练样本集$\{T_i \mid i=1,\cdots,M\}$，分类目标为测试集$\{T_i \mid i=M+1,\cdots,M+N\}$，通过已知分类的训练样本集基础上的有监督学习，建立基于 SVM 模型的分类机制对未知分类的测试集进行分类预测. 基于 SVM 的 fMRI 时间序列数据分类模型在解决小样本问题上，能够最大化利用原始脑区多维时间序列数据的特征属性信息，挖掘其有效模式与关系，达到分类、预测、识别的最佳效果.

针对包含 M 个样本的训练集，每个样本可表示为一个二元组(x_i, y_i)，其中 $\boldsymbol{x}_i=(x_{i1}, x_{i2}, \cdots, x_{id})^{\mathrm{T}}$ 对应第 i 个样本的 d 个属性集，$y_i\{-1,1\}$为类标号. 根据非线性 SVM 分类原理，将样本中特征属性集 x_i 通过非线性变换 $\Phi(\boldsymbol{x}_i)$映射到高维空间中，在变换空间求解最优分类面 $\boldsymbol{w}\cdot\Phi(\boldsymbol{x}_i)+b=0$，其中 $\boldsymbol{w}$ 表示权值矢量，b 为阈值，$\boldsymbol{w}$ 和 b 确定分类面位置. 对于过于接近超平面或者是被错误分类的向量 $\boldsymbol{x}_i$，引进松弛变量 ξ_i 和惩罚系数 c. 由此，基于训练样本集的分类问题可转化为以下优化问题：

$$\min\left(\|\boldsymbol{w}\|^2+\frac{c}{2}\sum_{i=1}^{m}\xi_i^2\right)$$
$$\text{s.t.},\ y_i(\boldsymbol{w}^{\mathrm{T}}\cdot\Phi(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i,$$
$$\xi_i\geqslant 0, i=1,\cdots,m \tag{4}$$

相应转化的对偶问题为

$$\max\left(\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\boldsymbol{\alpha}_i^{\mathrm{T}}y_iy_j\Phi(\boldsymbol{x}_i)^{\mathrm{T}}\Phi(\boldsymbol{x}_j)\boldsymbol{\alpha}_j\right)$$
$$\text{s.t.},\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha}=0, 0\leqslant\alpha_i\leqslant c, i=1,\cdots,m \tag{5}$$

基于拉格朗日乘子 α_i 的求解，可按式(6)—(7)计算参数 $\boldsymbol{w}$ 和 b：

$$\boldsymbol{w}=\sum_{i=1}^{m}\boldsymbol{\alpha}_iy_i\Phi(\boldsymbol{x}_i) \tag{6}$$

$$\alpha_i\left(y_i\left(\sum_{j=1}^{m}\alpha_jy_j\Phi(\boldsymbol{x}_j)\cdot\Phi(\boldsymbol{x}_i)+b\right)-1\right)=0 \tag{7}$$

引入满足 Mercer 条件的非线性核函数 $K(x_i, x_j)$表示变换空间中两向量的点积，即 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\Phi(\boldsymbol{x}_i)^{\mathrm{T}}\Phi(\boldsymbol{x}_j)$，由此确定针对分类测试集 z 的决策函数为

$$f(z)=\operatorname{sgn}\left(\sum_{i=1}^{m}\boldsymbol{\alpha}_iy_iK(\boldsymbol{x}_i,z)+b\right) \tag{8}$$

其中，构造支持向量 $\boldsymbol{x}_i$ 与输入空间抽取的向量 $\boldsymbol{x}_j$ 之间的内积核是构造支持向量机的关键，为获得更为理想的 fMRI 时间序列分类性能，选用以下高斯径向基核函数（Gaussian radial basis function, GRBF）：

$$K_{\mathrm{RBF}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-g\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2)\Phi(\boldsymbol{x}_i)^{\mathrm{T}}\Phi(\boldsymbol{x}_j) \tag{9}$$

4　基于 PSO 算法的分类模型参数优化

在面向 fMRI 时间序列数据的 SVM 分类模型中，支持向量机通过核函数将输入空间中线性不可分的数据映射到高维特征空间，从而使问题变得线性可分，因此 SVM 分类模型的参数优化选择对整体分类性能有着非常关键的作用. 决策惩罚参数 c 用来调节学习机置信范围和经验风险的比例，协调控制错分率和算法复杂度，使学习机具有最佳能力；而 RBF 核参数 g 表示 RBF 核的宽度，核参数的改变意味着非线性空间映射函数的改变，从而影响样本特征子空间分布的复杂程度. 为寻找最优的决策惩罚参数和核参数，构建基于 PSO 算法的参数优化模式，对于在大规模样本中寻找最佳参数，具有更广泛的适用性.

PSO 算法是一种基于群体智能的演化计算，每个解被称为“粒子(particle)”，所有粒子共存合作择优，每个粒子通过自身最优值和粒子群的最优值，向更好的位置“飞行”搜索最优解. 每个粒子都有一个适应值来衡量自身情况，由速度值决定其飞行的方向和距离，同时所有粒子都追随当前最优粒子在解空间中进行搜索. 设解空间为 D 维，群中粒子总数为 R，第 i 个粒子位置表示为向量 $\boldsymbol{X}_i=(x_i^1, x_i^2, \cdots, x_i^D)$，该粒子已搜索到最优位置为 $\boldsymbol{P}_i=(p_i^1, p_i^2, \cdots, p_i^D)$，粒子群已搜索到的最优位置为 $\boldsymbol{P}_g=(p_g^1, p_g^2, \cdots, p_g^D)$，第 i 个粒子的飞行速度为 $\boldsymbol{V}_i=(v_i^1, v_i^2, \cdots, v_i^D)$，则每个粒子的速度和位置按式(10)—(11)获得：

$$v_i^d(t+1)=\omega v_i^d(t)+c_1\eta_1\lfloor p_i^d(t)-x_i^d(t)\rfloor+ c_2\eta_2\lfloor p_g^d(t)-x_i^d(t)\rfloor \quad (10)$$

$$x_i^d(t+1)=x_i^d(t)+\beta v_i^d(t+1) \quad (11)$$

其中,$1\leqslant i\leqslant R$, $1\leqslant d\leqslant D$;c_1 与 c_2 为大于 0 的常数,即加速因子;η_1 与 η_2 为[0,1]之间的随机数;ω 为惯性系数,较大值意味着对解空间进行大范围探查,较小值意味着进行小范围探查;β 为约束因子,用于控制速度权重;第 d 维的位置变化范围为$[-x_{\max}^d, x_{\max}^d]$,速度变化范围为$[-v_{\max}^d, v_{\max}^d]$,即在迭代过程中若 v_i^d 和 x_i^d 超出边界值则将其设为边界值;适应度函数值为训练样本集进行交叉验证的准确率. 由此,针对 fMRI 时间序列 SVM 分类器的决策惩罚参数 c 和 RBF 核参数 g,按如下算法流程进行参数寻优:

(1)初始化,设定 PSO 参数初始值,即加速因子 c_1(局部搜索能力初始值)和 c_2(全局搜索能力初始值)、种群最大数量 R(最大进化数量)、速率、速率和参数关系、速率更新速度弹性系数 ω、种群更新速度弹性系数 β、决策惩罚参数 c 和核参数 g 的最大值与最小值.

(2)初始解生成,即产生初始粒子和速度,在 c 和 g 设定的初始范围内随机产生种群和速度的初始值,并将 c 和 g 的值输入 SVM 分类器,计算其交叉验证准确率作为适应度值,确定全局极值及其对应的全局极值点(即最优粒子),并计算每一代种群的平均适应度.

(3)迭代寻优直到最大迭代次数或满足迭代终止条件——根据粒子群速度更新式(10)计算每个粒子的速度,根据种群位置更新式(11)调整每个粒子的位置,经过自适应粒子变异后基于 SVM 分类模型计算其交叉验证准确率作为适应度值. 然后,进行判定,如果新位置的适应度优于局部最优粒子则用新粒子代替局部最优粒子,如果种群中的最优粒子优于全局最优粒子则将种群中的最优粒子替换全局最优粒子,并计算全局适应度和平均适应度.

(4)返回最优粒子和适应度值,得到最佳的(c, g)和交叉验证准确率,利用得到的最佳参数组合建立 SVM 分类模型进行训练并对分类测试集进行类别预测.

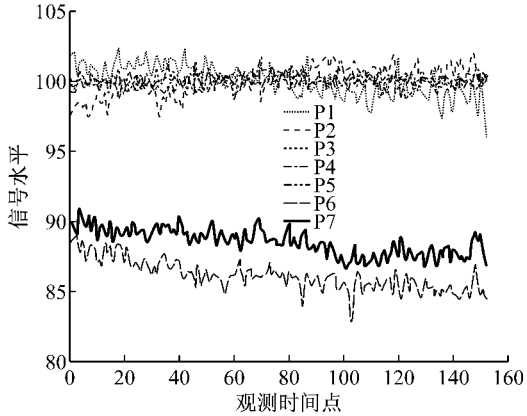
5 仿真实验与实证分析

5.1 数据来源

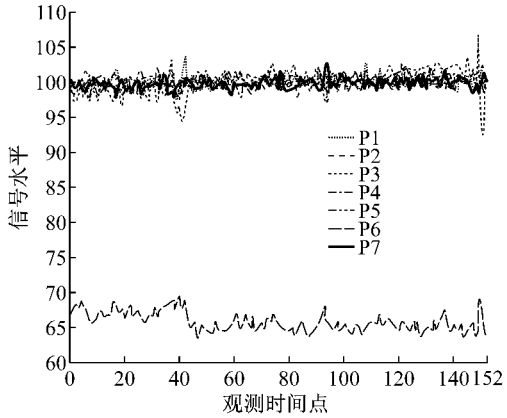
本文以精神抑郁症为对象进行实验仿真,实验数据由某知名大学附属医院精神医学实验室提供. 所采集的磁共振成像数据经过初期处理与分析,获得 fMRI 时间序列数据集. 所有被试者均符合相应医学标准,且抑郁症组和健康对照组在年龄、性别分布及受教育年限分布的差异无统计学意义. 一共获取 33 个抑郁症样本和 33 个健康样本,每个样本包括 7 个脑区在 7 分 32 秒内的 152 个观测值,采样间隔为 3 s,即对于每个样本均可由一个 7×152 的原始矩阵表示,66 个样本形成多维时间序列数据集. 实验中将每类样本的前 20 个样本作为训练集,另外 13 个样本作为测试集.

5.2 针对 fMRI 时间序列特征表示的相关实验

为了对原始 fMRI 时间序列数据具有直观的认识,图 1 给出原始数据集中一个健康样本和一个病人样本的信号水平对比图,每个样本均包含 7 个脑区的 152 个观测值. 从图 1 可看出,抑郁症样本和正常样本各脑区(P1~P7)时间序列可相对视为平稳序列或者去除趋势项后的平稳序列,围绕均值上下波动,在时间维度上存在着很明显的相似性,但两种类型的样本在 P6 和 P7 两个脑区的均值存在明显差异.



a 某抑郁症样本各脑区信号水平



b 某正常样本各脑区信号水平

图 1 随机抑郁症样本和正常样本原始信号水平对比

Fig. 1 Original signal level contrast between random samples of depression and normal samples

基于所构建 fMRI 脑区时间序列数据，利用自回归 AR 模型构建其相应的特征属性表示. 同时，为展现基于 AR 模型的特征提取模式有效性，在相同时间序列数据集上实现另一种比较广泛采用的时间序列数据特征提取方法，即基于主成分分析(principal component analysis，PCA)的表示. 有关 fMRI 时间序列数据特征提取的相关实验统计结果，见表 1. 表中，括号内的数字表示计算结果的由来，"/"表示相除. 其中，考虑到如果对特征样本数据进行归一化处理，将原始数据映射到同一范围内以在一定程度上简化计算提高训练速度，但归一化并非必须采用的预处理方法，如处理不当会影响分类效果，因此实验中包括对特征样本数据不作任何归一化处理、以及进行[0，1]和[－1，1]范围内的归一化处理两种不同方式.

表 1　fMRI 时间序列数据特征提取的实验结果

Tab.1　The experimental results of the data feature extraction of the fMRI time series

特征表示模式	处理方法	训练集分类准确率/%	测试集分类准确率/%	最佳参数选项	交叉验证准确率/%
基于 AR 的特征表示	不进行归一化	100(40/40)	80.77(21/26)	$c=0.71, g=1$	85.0
	[0，1]归一化	95(38/40)	80.77(21/26)	$c=0.71, g=22.63$	80.0
	[－1，1]归一化	95(38/40)	80.77(21/26)	$c=0.71, g=5.66$	80.0
基于 PCA 的特征表示	不进行归一化	90(36/40)	73.08(19/26)	$c=4, g=18$	85.0
	[0，1]归一化	92.5(37/40)	76.92(20/26)	$c=32, g=16$	82.5
	[－1，1]归一化	97.5(39/40)	76.92(20/26)	$c=11.31, g=11.31$	82.5

从表 1 中可看出，在使用相同参数设置和归一化预处理的情况下，基于 AR 模型和基于 PCA 方法的特征属性表示都能够获得较高的分类预测准确率，而基于 AR 模型的特征属性表示在不经归一化处理的情况下，其分类预测准确率和交叉验证准确率均为最高，且训练集能够被 100%正确分类，故不经归一化的基于 AR 模型的特征样本数据更有利于分类模型的训练，能够更有效地表达样本的特征属性. 同时，也表明是否需要进行归一化预处理往往因样本数据而异，选择哪种归一化方式并非必须. 因此，本文实验均以不经归一化的基于 AR 模型的特征数据作为样本集，对于每个样本最终都能够提取出 7 个属性值代替最初的 7 个多维时间序列. 有关基于 AR 模型的特征样本数据的 box 可视化图，如图 2 所示.

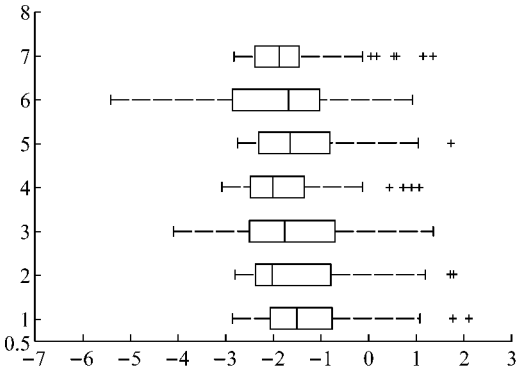


图 2　基于 AR 模型特征样本数据的 box 可视化图

Fig.2　The box visual figure basing on the characteristics sample data of the AR model

5.3　针对 fMRI 时间序列优化分类模型的相关实验

针对所构建的基于 PSO-SVM 的 fMRI 时间序列优化分类模型，在 PSO 算法优化参数选择方面，考虑到编码方式和适应度评价函数两个关键问题，编码方式选择采用实数编码，适应度函数值则取针对训练集进行交叉验证意义下的准确率. 其中，设定 PSO 参数局部搜索能力初始值 c_1 为 1.5，全局搜索能力初始值 c_2 为 1.7，最大进化数量初始值为 200，种群最大数量为 20，速率和参数关系为 0.6，速率更新速度弹性系数为 1，种群更新速度弹性系数为 1. 有关基于 SVM 模型的分类机制，其决策惩罚参数 c 的最大值设置为 2^5，最小值为 2^{-5}；核参数 g 的最大值设置为 2^5，最小值为 2^{-5}. 有关 fMRI 时间序列数据优化分类的相关实验统计结果，见表 2. 从表 2 中可看出，十次随机实验中平均预测准确率达到 82.69%，十次中仅一次相对稍低为 76.92%，且十次交叉验证准确率为 87.50%. 由此可观测到，基于 SVM 的分类模型在解决面向小样本 fMRI 时间序列数据的分类诊断问题上具有良好分类性能，且其惩罚参数和相关核函数参数的优选对于分类性能至关重要. 在同一样本数据及参数范围内，基于 PSO 算法参数寻优来搜索最佳参数的 SVM 分类性能表现良好. 基于 PSO-SVM 的 fMRI 时间序列数据分类诊断机制能够正确识别患病组和正常组，尽管训练样本并非大规模，测试集有噪声等不确定因素影响，依然取得 82.69%的平均分类准确率，再次表明该分类诊断模型在具体病症预测与识别中的实用性和有效性.

表2 fMRI时间序列优化分类模型的实验结果

Tab. 2 The experimental results of the optimization classification model of the fMRI time series

运行次数	最佳 c	最佳 g	预测准确率/%	交叉验证准确率/%
1	1.460 4	1.330 3	84.62	87.50
2	4.539 2	1.225 7	84.62	87.50
3	5.056 2	1.232 5	80.77	87.50
4	1.685 3	1.218 0	84.62	87.50
5	7.986 0	1.234 0	80.77	87.50
6	2.389 6	1.250 8	84.62	87.50
7	1.727 7	1.316 3	84.62	87.50
8	3.583 9	1.233 8	76.92	87.50
9	2.316 8	1.240 2	84.62	87.50
10	10.423 2	1.169 4	80.77	87.50
	平均值		82.69	87.50
	最大值		84.62	87.50
	最小值		76.92	87.50

另外,为进一步说明基于PSO算法的分类参数优化策略的优势,有关第7次运行中寻找最佳参数设置的适应度曲线如图3所示.从图3中可明显观察到,种群的平均适应度大都在81%以上,最佳适应度稳定在87.5%.在相同的参数范围内,基于PSO算法进行参数寻优的分类性能较为理想,能够找到更好的惩罚参数 c 和核参数 g 使分类准确率提高至84.62%,其相应平均最佳适应度提高2.94%,测试集分类准确率提高2.38%.因此,利用基于PSO算法的分类参数优化,最终可得到多组分类参数 c 和 g 的最优组合.但需要注意的是,由于惩罚因子 c 用于控制模型复杂度和逼近误差,其值越大则对数据的拟合程度越高,这将导致泛化能力降低,因此往往取 c 值最小且预测准确率最高的分类参数组合(参数 $c_1=1.5$, $c_2=1.7$,终止代数为200,种群数量 $P_{OD}=30$).

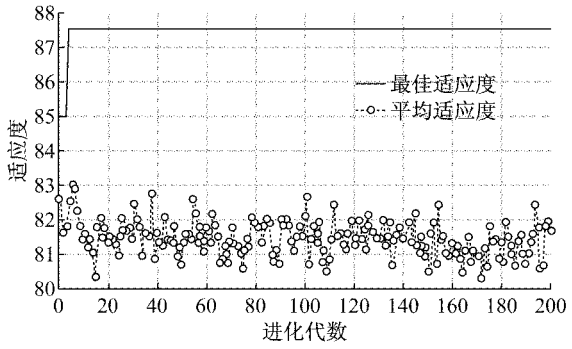


图3 PSO算法寻找最优参数的适应度曲线

Fig. 3 The fitness curves of the PSO algorithm

5.4 针对不同优化分类机制的对比实验

为进一步展现所提出基于PSO-SVM的fMRI时间序列数据分类诊断模型的优势,也开展有关针对不同优化分类机制的对比实验.融合网格搜索(grid search)与SVM的优化分类模型是一种应用比较广泛的分类机制,其参数优化方法建立在交叉验证的基础上,利用设定的参数最大值和最小值确定一个网格作为搜索范围,在该范围内按照初始设定的步长寻找使得交叉验证准确率最高的参数组合,确定最佳参数设置后即可对SVM分类模型进行训练和分类[11].从整体架构上来讲,该模型与本文所提出基于PSO-SVM的优化分类模型比较类似,有关两者分类性能的相关对比实验统计结果,见表3.

表3 两种不同优化分类模型的实验结果对比

Tab. 3 The experimental results of two different optimization classification model

优化分类模型	最佳参数	交叉验证准确率/%	预测准确率/%	健康组正确识别率/%	患者组正确识别率/%
融合网格搜索与SVM	$c=0.71, g=1$	85.00	80.77	76.92	84.62
基于PSO-SVM	$c=1.46, g=1.33$	87.50	84.62	84.62	84.62

从表3中可看出,基于PSO-SVM优化分类模型分类性能显著高于融合网格搜索与SVM的优化分类模型分类性能,交叉验证准确率从后者的85%提高至前者的87.50%.在融合网格搜索与SVM的优化分类模型中,利用参数寻优策略所获得最佳参数设置 $c=0.71, g=1$,基于整体测试集针对患者组和健康组的分类准确率为80.77%.在基于PSO-SVM的优化分类模型中,利用参数寻优策略所获得最佳参数设置 $c=1.46, g=1.33$,基于整体测试集针对患者组和健康组的分类准确率提高至84.62%,同融合网格搜索与SVM的优化分类模型相比较分类准确率提高2.38%,而通常现实生活中对于抑郁症的正确识别率只有70%左右.实验结果表明所提出的基于PSO-SVM优化分类模型比融合网格搜索与SVM的优化分类模型更胜一筹,同时也反映使用PSO算法进行SVM参数选优的可行性与合理性.基于PSO-SVM优化分类机制是针对fMRI时间序列数据与具体病症之间关联决策的较好途径,可有效支持这类特殊时间序列数据信息的准确分类判别.

6 结论

本文主要介绍在基于PSO-SVM的fMRI时间

序列分类诊断方面所做的一些研究工作，其中提出以基于 SVM 的分类模型为基本框架，同时结合基于 AR 模型的特征表示与基于 PSO 算法的参数优化处理来实现 fMRI 时间序列数据的自动预测与分类. 从目前的实验结果来看，已达到比较令人满意的效果. 该模型除表明统计学习的独特优势之外，同时也表明在统计学习方法基础上，适当应用时间序列数据相关的特有属性信息与启发式寻优策略，是提高 fMRI 时间序列分类诊断性能的有效方式. 目前，脑区多维时间序列样本集还有待于扩充，同时为更为丰富优化分类过程中所利用的特征属性信息，也要进一步深入考虑新建形式的特征模板来反映和揭示多脑区时间序列所包含的时空信息，在更大程度上提高模型化与关联度. 上述这些方面都将成为在未来研究工作中所要关注的问题.

参考文献：

[1] Guo H, Cao X H, Liu Z F, *et al*. Machine learning classifier using abnormal brain network topological metrics in major depressive disorder [J]. Neuroreport, 2012, 23(17): 1006.

[2] Klöppel S, Stonnington C M, Chu C, *et al*. Automatic classification of MR scans in Alzheimer's disease [J]. Brain, 2008, 131(3): 681.

[3] Li Y, Wang Y, Wu G, *et al*. Discriminant analysis of longitudinal cortical thickness changes in Ahheimer's diseaseusing dynamic and network features [J]. Neurobiology of Aging, 2012, 33(2): 427.e15.

[4] Kamitani Y, Tong F. Decoding the visual and subjective contents of the human brain [J]. Nature Neuroscience, 2005, 8(5): 679.

[5] Wee C Y, Yap P T, Zhang D Q, *et al*. Identification of MCI individuals using structural and functional connectivity networks [J]. NeuroImage, 2012, 59(3): 2045.

[6] Xie S Y, Guo R, Li N F, *et al*. Brain fMRI processing and classification based on combination of PCA and SVM [C]// Proceeding of the 2009 International Joint Conference on Neural Networks (IJCNN'09). Atlanta: IEEE, 2009: 3384-3389.

[7] Norman K A, Polyn S M, Detre G J, *et al*. Beyond mind-reading: multi-voxel pattern analysis of fMRI data [J]. Trends in Cognitive Sciences. 2006, 10(9): 424.

[8] Fu C H, Mourao-Miranda J, Costafreda S G, *et al*. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression [J]. Biological psychiatry, 2008, 63(7): 656.

[9] Chupin M, Gerardin E, Cuingnet R, *et al*. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI [J]. Hippocampus, 2009, 19(6): 579.

[10] Gong Q, Wu Q, Scarpazza C, *et al*. Prognostic prediction of therapeutic response in depression using high-field MR imaging [J]. Neuroimage, 2011, 55(4): 1497.

[11] 相洁，陈俊杰. 基于 SVM 的 fMRI 数据分类：一种解码思维的方法 [J]. 计算机研究与发展，2010，47(2)：286.
XIANG Jie, CHEN Junjie. SVM based fMRI data classification: an approach to decode mental state [J]. Journal of Computer Research and Development, 2010, 47(2): 286.

[12] Banerjee T P, Das S. Multi-sensor data fusion using support vector machine for motor fault detection [J]. Information Sciences, 2012, 217: 96.

[13] de Moraes R M, dos Santos Machado L. Online assessment in medical simulators based on virtual reality using fuzzy Gaussian naive bayes [J]. Multiple-Valued Logic and Soft Computing, 2012,18(5-6): 479.

[14] Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMPJ: a tutorial overview [J]. Neuroimage, 2009, 45(1): 199.

[15] Feis D L, Brodersen Kay H, von Cramon D Y, *et al*. Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data [J]. Neuroimage, 2013, 70: 250.

[16] Mueller S G, Young K, Hartig M, *et al*. A two-level multimodality imaging Bayesian network approach for classification of partial epilepsy: preliminary data [J]. Neuroimage, 2013, 71: 224.

[17] Calhoun V D, Maciejewski P K, Pearlson G D, *et al*. Temporal lobe and "default" hemodynamic brain modes discriminate between schizophrenia and bipolar disorder [J]. Hum Brain Mapp, 2008, 29(11): 1265.

[18] Davatzikos C, Shen D, Gur R C, *et al*. Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities [J]. Arch Gen Psychiatry, 2005, 62(11): 1218.