doi:10.6043/j.issn.0438-0479.201606009

**基于PCA和LDA融合的阿尔茨海默病分类算法**

杨晨晖，余传健

（厦门大学信息科学与技术学院，福建 厦门 361005）

**摘要**：在阿尔茨海默病( Alzheimer's disease，AD)诊断方法中，通过对脑图像分析已成为准确诊断的一种重要手段。针对从单一脑图像模态MRI中提取的特征,提出了一种基于主成分分析（PCA）和线性鉴别分析（LDA）融合的AD分类识别算法。该方法首先对从磁共振图像(MRI)图像中获取的特征进行主成分分析，其次对低维的特征进行线性鉴别分析获取组合特征向量，最后采用最邻近算法，利用获取的组合特征向量对未知状态类型进行分类识别。实验表明，该算法与其他相关算法相比，具有较高的识别准确率、敏感性、特异性，这说明了算法的有效性。

**关键词：**阿尔茨海默病；脑图像分析；主成分分析；线性鉴别分析；最邻近算法

**中图分类号：**TP399 **文献标识码：**A

阿尔茨海默病( Alzheimer's disease，AD) 又称原发性老年痴呆症，是致死性神经退行性疾病之一，主要表现为认知和记忆能力降低。阿尔茨海默病由于发病时间较晚，而且不显示特异性指标，在发病早期很容易被忽视，而当被确诊时已经到了晚期，错过最佳的治疗时期。因此，对阿尔茨海默症的早期诊断研究是非常重要的。

近期，随着计算机学科的发展，阿尔茨海默病相关研究工作得到了巨大的发展，许多机器学习和模式分类方法广泛应用在AD/MCI医学图像分析。早期的阿尔茨海默病研究主要集中在从单一的医学图像模态，如磁共振图像(MRI)和氟脱氧葡萄糖-正电子发射断层扫描（FDG-PET），提取特征(基于感兴趣区域(ROI)或体素)进行疾病分类。2008年，He等首次利用MRI计算出AD患者的大脑皮层厚度，从而构建了他们的脑部结构网络。从脑连接科学的角度经过对该网络的定性、定量分析，发现它们与正常对照组相比，全脑的结构与功能、大脑的局部脑区、脑区与脑区之间的连接等均存在异常。同年，Querbes等利用计算出的大脑皮层厚度，以单个脑图像个体为单位建立大脑皮层厚度图谱，从图谱提取厚度特征进行分类。2014年，杨文璐等利用MRI分割出海马体，再利用独立成分分析提取海马体表面形态学特征实现阿尔茨海默症的分类。近年来，随着越来越多的示踪剂被发现，PET图像逐渐显示出在AD诊断预测方面的优势。Silveira计算FDG-PET的体素值，使用boosting算法得出了90.97%的AD诊断率和79.63% 的MCI诊断率。Shen将PET图像视为一个矩阵，使用PCA降维方法做特征选择，不考虑图像本身的医学信息，在2015年的大脑图像处理顶级会议IPMI上发表。

阿尔茨海默病的发展过程会经过三种状态：正常（Normal Control,NC)、轻度认知障碍（Mild Cognitive Impairment,MCI)、阿尔茨海默病( Alzheimer's disease，AD)。另外，轻度认知障碍在一段时间后可能转换成正常状态，也可能转换成阿尔茨海默病，所以MCI可以分为MCI-NC（转换成正常）、MCI-C（转换成阿尔茨海默病）。在本文的实验，我们采用四种状态：NC、AD、MCI-NC、MCI-C，后面两种状态统称为MCI。

**1 基于PCA的分类算法**

PCA (Principal Component Analysis)即主成分分析技术,旨在利用降维的思想，将原始高维空间样本投影至低维空间。在低维空间上上，样本的采样密度增大，利于分类，而且还有最少的信息损失量。具体分类算法步骤描述如下：

（1）计算所有训练样本的平均向量：



其中，M为训练样本集中样本的个数，为第i个训练样本，，d表示样本的维度。

（2）计算差值向量。每个训练样本与样本平均向量的差值向量为：



（3）训练样本的协方差矩阵可以表示为：



上式中A= []。

（4）计算训练集协方差矩阵C的特征值以及特征向量。采用能量维数法来选取最优特征向量，选择最大的前k个特征值对应的特征向量作为最优特征向量，使得：



其中为自定义最佳能量阈值，在实验中我们取。

（5）根据步骤（4）选取的k个特征向量组成子特征空间中的特征向量矩阵

（6）计算训练样本X的降维向量矩阵Q。特征向量矩阵U可以看作是训练特征在子特征空间的投影向量矩阵，每个训练样本在子特征空间的投影特征为：



为训练样本的降维向量矩阵，每列对应X中的一个训练样本，。

（7）将需要分类的样本特征减去训练样本平均特征向量，得到一个差值向量r；

（8）将差值向量r做线性变换，获得测试样本特征的最佳分类特征向量：



（9）采用最近邻分类算法分类：计算最佳分类特征和每个训练样本的最佳分类特征的欧几里德距离，标记最小的距离，则待识别特征样本属于类别j。

上述步骤（1）-（6）属于训练步骤，步骤（7）-（9）属于分类步骤。

**2 基于LDA的AD分类算法**

 LDA (Linear Determinant Analysis)即线性鉴别分析，其基本思想是将高维的模式样本投影到最佳鉴别矢量空间，以达到抽取分类信息和压缩特征空间维数的效果。LDA采用了样本的类标作为先验知识，将样本特征向量投影到某几个最佳的投影直线，使得投影后向量的类间离散度矩阵和类内离散度矩阵的比率最大化，从而可以提高识别率。总之，LDA可以找到一组特征子维度，在这些维度上的投影产生的特征有类内距离最小，类间距离最大。

（1）计算类间离散度。训练特征集X的类间离散度表示为：

 

其中，P(i)为第i个训练特征样本的先验概率，其中为所有特征样本向量的均值样本特征，表示为第 i 个训练样本特征。

（2）计算类内离散度。训练特征集的类内离散度表示为：

 

其中，M为训练样本个数，为第k个类别标记。k = 1,2,...,K，K为类别数。

（3）计算特征向量矩阵V。求解矩阵的特征值，选取前t个最大特征值对应的特征向

量，构成子特征空间的特征向量矩阵。

（4）计算训练样本X的降维向量P。特征向量矩阵V可以看作是训练样本在子特征空间的投影向量矩阵，每个训练样本在子特征空间的投影向量为：



为训练样本的降维向量矩阵，每列对应X中的一个训练样本，。

（5）将需要分类的样本特征减去训练样本平均特征向量，得到一个差值向量r；

（6）将差值向量r做线性变换，获得测试样本特征的最佳分类特征：



（7）采用最近邻分类算法分类：计算最佳分类特征和每个训练样本的最佳分类特征的欧几里德距离，标记最小的距离（i,j = 1,2,...,C），则待识别样本属于类别j。

上述步骤（1）-（4）属于训练步骤，步骤（5）-（7）属于分类步骤。

**3 PCA和LDA融合的AD分类算法**

考虑到PCA是无监督的学习方法，它仅仅只是把整体数据映射到最方便表示这组数据的坐标轴上，映射没有利用任何数据内部的分类信息。虽然用主要的特征代替其他相关的非主要的特征，数据特征在表示上更加方便，数据信息损失最少，但是在分类上也许会更加困难。LDA是有监督的降维学习方法，由于增加了数据本身的性质，所以数据有聚集的映射到某几个坐标轴上，再取这几个坐标轴上的投影作为新的特征，从而使得数据更容易分类。如图1所示二维实例，左边表示PCA降维，右边表示LDA降维。可以看出，PCA降维之后，数据依然密集，不容易区分；LDA降维之后，数据的区分度比较大，数据容易区分，准确度比较高。



（a）PCA (b) LDA

图 1 2D数据降维

Fig.1 Reduction in Two Dimension Data

综合上面主成分分析和线性判别分析方法的优点和缺点，本文分类算法主要利用PCA的无监督降维和LDA的有监督降维的优缺点互补。算法实现主要步骤描述如下：

1. 在最原始训练特征上利用PCA无监督的降维，计算特征向量矩阵，再得到特征子空间的降维向量矩阵。其中，M为训练样本个数，的维度为k。
2. 在矩阵上利用LDA进行有监督的降维，计算特征向量矩阵，再得到最佳分类特征空间的降维向量矩阵,的维度为t（t<=k）。
3. 将需要分类的样本特征减去训练样本平均特征向量，得到一个差值向量r；
4. 将差值向量r做线性变换，获得测试样本特征的最佳分类特征：



（5）采用最近邻分类算法分类：计算最佳分类特征和每个训练样本的最佳分类特征的欧几里德距离，标记最小的距离（i,j = 1,2,...,C），则待识别样本属于类别j。

上述步骤（1）-（2）属于训练步骤，步骤（3）-（5）属于分类步骤。

**4 结果分析**

 本文在实验中所采用的数据集，包含727张图片（其中94张AD、296张NC、309张MCI-NC、120张MCI-C图片），是可以从阿尔茨海默症神经影像（Alzheimer’s Disease Neuroimaging Initiative,ADNI）数据库（http://adni.loni.usc.edu/）中下载的MRI医学图像。每幅图像通过一系列的预处理，包括头洞校正、头骨剥离、ROI提取等，我们得到了93维的特征。在本文中，我们实现了三组数据实验，NC与AD、MCI-NC与MCI-C、MCI与NC对比，由于数据集样本比较小，我们采用十折交叉验证进行结果测试。为了进一步降低偶然性，采用运算100次，最终结果取100次结果中最差的一次。

本文算法的实现运行环境为：Windows 7，i5-4590处理器，主频3.30GHz，8G内存。

为了测试PCA和LDA融合算法的优越性，首先，本文将文献的算法结果与PCA和LDA融合算法实现得到的分类结果各项评价指标进行对比，实验结果如表1所示。实验采用ACC、SEN、SPE三个评判标准，计算公式如下所示：

Accuracy(ACC) =（TP+TN）/(TP+TN+FP+FN)

Sensitivity(SEN) = TP/(TP+FN)

 Specificity(SPE) = TN/(TN+FP)

其中TP表示真正类：正类预测成正类的个数，TN表示真负类：预测负类成负类的个数， FP表示假正类：负类预测成正类的个数，FN表示假负类：正类预测成负类的个数。

其次，横向的验证我们算法的有效性。因为我们提出的算法是PCA和LDA融合实现的算法，所以本文算法和单纯的PCA、LDA算法结果比较，实验结果如表2所示。

表1 不同文献算法结果对比

Tab.1 The classification performance with different literature algorithm

|  |  |  |  |
| --- | --- | --- | --- |
| 实验方法 | AD vs NC | MCI-C vs MCI-NC | NC vs MCI |
| ACC(%) | SEN(%) | SPE(%) | ACC(%) | SEN(%) | SPE(%) | ACC(%) | SEN(%) | SPE(%) |
| 文献[22] | 86.2 | 86 | 86.3 | 55.3 | 52.7 | 53.2 | 72 | 78.5 | 59.6 |
| 文献[23] | 85.7 | 85.1 | 86.8 | 54.9 | 50.4 | 56.0 | 70.6 | 72.4 | 60.1 |
| 文献[24] | 88.68 | 84.51 | 92.5 | 52.22 | 44.19 | 58.21 | 73.12 | 78.28 | 63.65 |
| 本文 | 94.5 | 94.9 | 92.9 | 70.2 | 65.4 | 71.3 | 76.8 | 70.3 | 89.8 |

表2 PCA、LDA、PCA和LDA融合算法实验结果对比

Tab.2 The classification performance compare fusion algorithm with baseline algorithm

|  |  |  |  |
| --- | --- | --- | --- |
| 实验方法 | AD vs NC | MCI-C vs MCI-NC | NC vs MCI |
| ACC(%) | SEN(%) | SPE(%) | ACC(%) | SEN(%) | SPE(%) | ACC(%) | SEN(%) | SPE(%) |
| PCA | 86.9 | 85.6 | 87.0 | 50.2 | 47.6 | 56.5 | 70.9 | 69.7 | 65.3 |
| LDA | 87.8 | 86.0 | 86.4 | 51.6 | 47.4 | 57.0 | 70.5 | 69.4 | 66.4 |
| PCA+LDA | 94.5 | 94.9 | 92.9 | 70.2 | 65.4 | 71.3 | 76.8 | 70.3 | 89.8 |

从表1可以，和其他单模态MRI图像算法比较，我们的算法在三组实验都获得比较好的结果，准确率分别为94.5%，70.2%，76.8%。在敏感度和特异性等重要评价指标上都比其他算法好，这验证了我们算法的有效性。从表2可以看出，融合的算法比单独的算法也体现出自己的优越性，这也可以从三组不同实验得出的准确性、敏感度、特异性对比中总结出来。

**5 总 结**

本文通过对阿尔茨海默病的MRI医学图像分析，提出了基于PCA和LDA融合的分类算法。实验部分通过三组实验：AD与NC、MCI与NC、MCI-C与MCI-NC，将我们提出的算法和单纯的PCA、LDA算法比较以及与现在比较好效果的算法结果进行比较。通过对比实验，结果显示PCA-LDA融合的算法有更成功的识别准确率和相对更好的鲁棒性，这对老年痴呆症的诊断具有重要的意义。将来研究改进的方向是怎样更加有效地融合PCA和LDA以及在分类算法上的改进。另外，如果数据条件允许，将会采用多模态数据进行疾病的预测识别。

**参考文献:**

[1]章莹,付伟.国外AD预防研究现状[J].实用老年医学,2014,28(8):685-687．

[2]Chincarini A, Bosco P, Calvini P, et al. Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease[J]. NeuroImage, 2011, 58(2): 469-480.

[3]Fan Y, Resnick S M, Wu X, et al. Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study[J]. Neuroimage, 2008, 41(2): 277-285.

[4]Liu M, Zhang D, Shen D, et al. Ensemble sparse classification of Alzheimer's disease[J]. NeuroImage, 2012, 60(2): 1106-1116.

[5]Drzezga A, Lautenschlager N, Siebner H, et al. Cerebral metabolic changes accompanying conversion of mild cognitive impairment into Alzheimer's disease: a PET follow-up study[J]. European journal of nuclear medicine and molecular imaging, 2003, 30(8): 1104-1113.

[6]Foster N L, Heidebrink J L, Clark C M, et al. FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease[J]. Brain, 2007, 130(10): 2616-2635.

[7]He Y, Chen Z, Evans A. Structural insights into aberrant topological patterns of large-scale cortical networks in Alzheimer's disease[J]. The Journal of neuroscience, 2008, 28(18): 4756-4766.

[8]Querbes O, Aubry F, Pariente J, et al. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve[J]. Brain, 2009, 132(8): 2036-2047.

[9]杨文璐，张小慢,刘凤侠.阿尔茨海默病患者海马体表面形态学特征提取[J].安徽大学学报(自然科学版),2014,38(5)：91-97.

[10]Silveira M, Marques J. Boosting Alzheimer disease diagnosis using PET images[C]// Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, 2010: 2556-2559.

[11]Shen D, Zhu H. Spatially Weighted Principal Component Regression for High-Dimensional Prediction[C]//Information Processing in Medical Imaging. Springer International Publishing, 2015: 758-769.

[12]周志华.机器学习[M].北京：清华大学出版社，2016：229-232.

[13]Scholkopft B, Mullert K R. Fisher discriminant analysis with kernels[J]. Neural networks for signal processing IX, 1999, 1(1): 1.

[14]Zhang D, Wang Y, Zhou L, et al. Multimodal classification of Alzheimer's disease and mild cognitive impairment[J]. Neuroimage, 2011, 55(3): 856-867.

[15]Suk H I, Shen D. Deep learning-based feature representation for AD/MCI classification [M]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer Berlin Heidelberg, 2013: 583-590.

[16]Jie B, Zhang D, Cheng B, et al. Manifold regularized multitask feature learning for multimodality disease classification[J]. Human brain mapping, 2015, 36(2): 489-507.

**Alzheimer's Disease Classification Algorithm Based on Fusion PCA and LDA**

 YANG Chenhui, YU Chuanjian

(School of Information Science and Engineering, Xiamen University,Xiamen 361005, China）

**Abstract:** In Alzheimer's Disease diagnostic method, analyzing brain image has become an important means to diagnosis accurately. In this papers, we only consider brain feature extracted from single modality brain images MRI, proposes a new identification algorithm based on principal component analysis (PCA) and Linear Discriminant analysis LDA for AD classification recognition algorithm.Firstly, it makes principal component analysis (PCA) in the original brain feature extracted from single modality brain images MRI;Secondly, it makes linear discriminant analysis (LDA) in low dimensional feature from PCA method so as to get the combinational feature vector;Finally,we adopt nearest neighbor algorithm incombinational feature vector toclassification and recognize unknown type brain features. Experimental results have shown that compared with the related algorithms,the proposed algorithm has higher classification Accuracy,Sensitivity and Specificity, so our algorithm is effective.

**Key words:** Disease diagnosis; Brain image analysis; Principal component analysis; Linear Discriminant Analysis; Nearest neighboralgorithm