

基于联合嵌入和稀疏学习的帕金森疾病分类

摘要

帕金森疾病是一种主要影响病人的运动神经系统的不可逆的神经衰退疾病。早期对帕金森病进行分类是至关重要的，可以从其发病初期开始减缓这种退化过程。复现目标提出了一种新颖的自适应无监督特征选择方法，通过利用纵向多模态数据的流形学习，促进早期帕金森病的诊断。具体而言，所提出的方法执行联合嵌入和稀疏回归，可以自适应地确定相似性矩阵和具有区分性的特征。同时，复现目标约束了受试者之间的相似性矩阵，并利用 $l_{2,p}$ 范数进行稀疏自适应控制，以获得多模态数据结构的内在信息。提出了一种有效的迭代优化算法来解决这个问题。复现目标在帕金森病进展标志倡议 (PPMI) 数据集上进行了大量实验，以验证所提出方法的有效性。结果表明，复现目标的方法提升了纵向数据分类和临床评分回归的性能，并超过了现有的最先进方法。此复现工作尝试复现了复现目标的特征选取过程，并对此过程进行了验证。

关键词：分类；嵌入学习；帕金森疾病；稀疏回归

1 引言

帕金森病 (PD) 是一种常见的老年神经退行性疾病。早期诊断对于减缓这种退化过程至关重要。多模态数据能够为神经退行性疾病的分类提供互补信息。例如，磁共振成像 (MRI) 中的灰质 (GM) 被广泛应用于获取神经细胞变化的信息。扩散张量成像 (DTI) 的第一特征值 (L1) 和第一特征向量 (V1) 分别表示最大的扩散系数和其方向向量。因此，L1 和 V1 可能对大脑中的神经退行性变化更为敏感。在使用多种模态的数据时，特征选择是发现与疾病相关特征的优秀手段，可以利用监督或非监督方法进行挖掘。此次选题的复现目标使用无监督学习，提取多模态数据中的特征，更好地利用了多模态的数据。选取此复现目标可以学习到过滤出对于特定任务最具信息量的特征，同时减少数据的复杂性和噪音，提高模型的泛化能力和性能。

2 相关工作

有许多监督式的联合多任务和无监督的嵌入方法。例如，基于 $L_{2,1}$ 范数的多模态多任务 (M3T) [3] 方法学习了一个特征选择模型，从每种模态中获取多个任务的共同相关特征。多模态稀疏学习 (MMSL) [2] 方法同时进行帕金森疾病的分类和回归预测，基于一个统一的多任务特征选择函数，考虑了响应矩阵中行和列之间的差异相似性。在 [1] 中，利用多模态数据通过关系正则化和判别学习来提高帕金森疾病的分类和回归预测性能。这些现有方法存在一

些局限性。例如，监督式多任务方法忽略了学习数据内在的结构信息。先前的无监督嵌入方法也可能由于原始特征空间中的噪音而计算出不准确的相似性矩阵。

3 本文方法

3.1 本文方法概述

此工作的输入为待分类样本数据矩阵 \mathbf{X} 。此工作处理样本相似性时有样本相似矩阵 \mathbf{S} 。在处理特征相似性时有特征相似矩阵 \mathbf{G} 。为使得 \mathbf{S} 有正确的邻居分布，此工作引入了 \mathbf{Q} 。为减少原高维数据中的噪声和重复，此工作同时采用了稀疏学习和嵌入学习，引入了特征权重矩阵 \mathbf{W} 。引入的矩阵组合计算构成了优化目标，方法后续为循环优化过程。优化过程首先初始化 \mathbf{S} 与 \mathbf{G} ，第二步利用 \mathbf{S} 和 \mathbf{G} 更新 \mathbf{Q} 与 \mathbf{W} 。第三步利用 \mathbf{Q} 与 \mathbf{W} 更新 \mathbf{S} 和 \mathbf{G} 。循环优化后得到特征权重矩阵 \mathbf{W} ，根据 \mathbf{W} 选取特征。

3.2 优化目标

复现目标的优化目标由提出的四个矩阵 \mathbf{S} 、 \mathbf{G} 、 \mathbf{Q} 、 \mathbf{W} 构成。优化目标可见公式 (1)。其中 \mathbf{x}^i 表示矩阵 \mathbf{X} 的第 i 行。 \mathbf{x}_j 表示第 j 列。 \mathbf{X} 的 $l_{2,p}$ 范数表示为 $\|\mathbf{X}\|_{2,p} = (\sum_i \|\mathbf{x}^i\|_2^p)^{\frac{1}{p}}$ 。 $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{G} \in \mathbb{R}^{d \times d}$, $\mathbf{Q} \in \mathbb{R}^{n \times r}$, $\mathbf{W} \in \mathbb{R}^{d \times m}$ 。其中 n 代表样本数， d 代表原数据维度数， r 代表分类任务分类数， m 代表 \mathbf{W} 的映射维度。 μ 参数和 θ 参数是防止无效解的正则参数。 λ 参数作用于特征权重矩阵 \mathbf{W} ， λ 增大时， \mathbf{W} 对应的特征选择更加稀疏。参数 σ 是确保样本相似矩阵 \mathbf{S} 能有正确邻居分布的参数，此参数在循环优化中变化来确保正确的邻居分布。参数 p 是 $l_{2,p}$ 范数的一部分。其中参数 m 、参数 λ 和参数 p 需要选取适当范围后遍历寻找最优。

$$\begin{aligned} & \sigma \text{tr}(\mathbf{Q}^T \mathbf{L}_S \mathbf{Q}) + \min \sum_{i,j} (\|\mathbf{x}^i \mathbf{W} - \mathbf{x}^j \mathbf{W}\|_2^2 s_{ij} + \mu s_{ij}^2 + \|\mathbf{w}^i - \mathbf{w}^j\|_2^2 g_{ij} + \theta g_{ij}^2) + \lambda \|\mathbf{W}\|_{2,p}^p \\ & \text{s.t. } \mathbf{s}^i \mathbf{1} = 1, 0 \leq s_{ij} \leq 1, \mathbf{g}^i \mathbf{1} = 1, 0 \leq g_{ij} \leq 1 \\ & \mathbf{Q} \in \mathbb{R}^{n \times r}, \mathbf{Q}^T \mathbf{Q} = \mathbf{E}, \mathbf{W}^T \mathbf{W} = \mathbf{E}. \end{aligned} \quad (1)$$

4 复现细节

此次复现实现了复现目标的特征选取过程。并使用 SVM 试验特征选取是否有效。关于代码结构，复现的特征选取过程内容位于 `getFeatureSelectionWeight.m` 中。对于 \mathbf{W} 的优化循环内容位于 `InterationW_impro_FS.m` 中。

4.1 与已有开源代码对比

没有参考相关源代码。

5 实验结果分析

5.1 实验设置

进行了三组实验，第一组实验是未进行特征选取。第二组实验是进行 PCA 特征选取。第三组实验是进行此复现的特征选取。对分类是否患有帕金森疾病任务，比较了三种方法的准确率和 F1-score。

5.2 实验结果及分析

三种方法的分类效能可见表1

表 1. 对于分类是否患有帕金森疾病的三种方法的效能

	ACC(%)	F1-score(%)
不使用特征选取	71.43±8.25	61.36±17.14
使用 PCA	78.45±6.89	63.99±10.54
使用复现方法	80.45±6.27	66.87±9.43

复现目标的论文描述的分类效能可见表2。

表 2. 复现目标对于分类是否患有帕金森疾病的二种方法的效能

	ACC(%)	F1-score(%)
使用 PCA	78.45±6.89	63.99±10.54
复现目标	83.33±7.92	71.17±13.00

可以看出复现的特征选取过程相比 PCA 有提升，但未达到复现目标论文的描述水平。

6 总结与展望

本次复现基本实现了复现目标的特征选取过程。但最后结果未达到论文描述。可能需要调整参数选取和循环次数。进行了论文描述的部分实验，未验证所有实验。

参考文献

- [1] Baiying Lei, Peng Yang, Tianfu Wang, Siping Chen, and Dong Ni. Relational-regularized discriminative sparse learning for alzheimer' s disease diagnosis. *IEEE transactions on cybernetics*, 47(4):1102–1113, 2017.
- [2] Haijun Lei, Zhongwei Huang, Jian Zhang, Zhang Yang, Ee-Leng Tan, Feng Zhou, and Baiying Lei. Joint detection and clinical score prediction in parkinson's disease via multi-modal sparse learning. *Expert Systems with Applications*, 80:284–296, 2017.

- [3] Daoqiang Zhang, Dinggang Shen, Alzheimer’s Disease Neuroimaging Initiative, et al. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer’s disease. *NeuroImage*, 59(2):895–907, 2012.