基于跨视角相似度顺序保持的基因特征提取方法

苏树智*①② 张开宇① 王子莹① 张茂岩①
①(安徽理工大学计算机科学与工程学院 淮南 232001)
②(合肥综合性国家科学中心能源研究院(安徽省能源实验室) 合肥 230031)

摘 要：基因表达数据通常具有维数高、样本少、类别分布不均等特点，如何提取基因表达数据的有效特征是基因分类研究的关键问题。该文借助相关分析理论，构建鉴别敏感的视角内相似度顺序保持散布并且约束鉴别敏感的视角间相似度相关，从而形成了一种新的基因特征提取方法，即相似度顺序保持跨视角相关分析(SOPACA)。该文方法在保持不同视角间特征类内聚集性和相似度顺序的同时具有较大的类间离散性。在癌症基因表达数据集上的良好实验结果显示了该文方法的有效性。

关键词：基因特征提取；相关分析理论；相似度顺序保持；鉴别敏感；癌症诊断
中图分类号：TN911.73; TP391.4 **文献标识码**：A **文章编号**：1009-5896(2023)01-0317-08
DOI: 10.11999/JEIT211126

A Gene Feature Extraction Method Based on Across-view Similarity Order Preserving

SU Shuzhi①② ZHANG Kaiyu① WANG Ziying① ZHANG Maoyan①
①(*School of Computer Science and Engineering, Anhui University of Science & Technology, Huainan* 232001, *China*)
②(*Institute of Energy, Hefei Comprehensive National Science Center, Hefei* 230031, *China*)

Abstract: Gene expression data is usually characterized by high dimension, few samples and uneven classification distribution. How to extract the effective features of gene expression data is a critical problem of gene classification. With the help of correlation analysis theory, the within-view and between-view discrimination sensitive similarity order scatter can be construsted, thus forming a new method of gene feature extraction, namely, Similarity Order Preserving Across-view Correlation Analysis(SOPACA). The proposed method not only maintains the intra-class aggregation and similarity order of features between different views, but also has a large distance between classes. Good experimental results on cancer gene expression datasets demonstrate the effectiveness of the method.

Key words: Gene feature extraction; Correlation analysis theory; Similarity order preserving; Discrimination sensitive; Cancer diagnosis

收稿日期：2021-10-14；改回日期：2022-01-10；网络出版：2022-02-02
*通信作者： 苏树智 sushuzhi@foxmail.com
基金项目：国家自然科学基金(61806006)，中国博士后科学基金(2019M660149)，合肥综合性国家科学中心能源研究院项目(19KZS203)，安徽省重点研发计划国际科技合作专项(202004b11020029)
Foundation Items: The National Natural Science Foundation of China (61806006), China Postdoctoral Science Foundation (2019M660149), The Project of Institute of Energy, Hefei Comprehensive National Science Center (19KZS203), The International Science and Technology Cooperation Project of Key Research and Development Plan in Anhui Province (202004b11020029)

1 引言

随着生物信息学的快速发展，人们对于癌症的研究已经发展到分子水平，脱氧核糖核酸(Deoxyribo-Nucleic Acid, DNA)微阵列技术[1]为人类在分子水平进行疾病诊断和治疗提供了全新手段，DNA微阵列技术可以大规模地快速检测基因表达情况获得基因表达数据，通过对基因表达数据进行分析可以了解细胞当前的生理状态区分癌变细胞与正常细胞，以便做出精准的诊断。基因表达数据样本个数通常为几十到几百，而每个样本的基因数量却成千上万，高维小样本作为基因表达数据的显著特性给大多数统计方法带来了挑战，对基因样本直接进行分类会存在维数灾难[2]问题，往往需要对基因表达数据进行维数约减，其目的是将原始数据投影到低维子空间以获得新的特征，该数据可以消除噪声和冗余信息利于后续处理。特征提取[3]作为最重要的维数约减方法之一，可以获得具有鉴别能力的特征，因此如何对关键有效基因进行特征提取成为基因分类研究的关键问题。

作为经典的单视角特征提取方法，主成分分析(Principal Components Analysis, PCA)[4]和线性判别分析(Linear Discriminant Analysis, LDA)[5]已经广泛应用于基因数据分析领域，Nakayama等人[6]使用基于高斯核的主成分分析方法用于基因表达数据聚类，讨论核参数的选择对于聚类性能的影响。Clayman等人[7]将PCA应用于研究DNA微阵列数据和临床变量之间的相互关系。Wang等人[8]提出了一种稀疏线性判别分析特征选择方法，与其他方法相比能够使用更少的特征或基因数量在降低错误分类率的情况下获得更好的结果。Lin等人[9]利用线性判别分析方法构造了阿尔茨海默病诊断框架，与之前研究中的方法相比，所提出的框架能够取得更好的分类性能。

随着信息的爆炸式增长，这种对于同一目标的一种表示的单视角学习方法已经不能满足研究者的需要，针对同一目标多种表示的多视角学习方法成为大势所趋，多视角学习既可以充分利用视角间的互补性，又能有效剔除视角间的冗余性，从而提取更具鉴别性的特征表示，在联合维数约减任务中，多视角数据可以发挥出比单视角数据更佳的识别性能。作为多视角学习的经典工具，典型相关分析(Canonical Correlation Analysis, CCA)[10]能够揭示两个不同视角之间的多元关系，CCA旨在找到一组基向量对，最大化从同一目标的两种不同视角获得的两个不同样本集之间的相关性，CCA已广泛应用于生物信息学领域，Lin等人[11]提出了组稀疏典型相关分析方法，引入组约束利用相关分析中的结构信息研究单核苷酸多态性与功能性磁共振成像测量的大脑活动之间的对应关系。Tenenhaus等人[12]提出了核广义典型相关分析方法，并提供了一个考虑块之间先验连接图的多块数据分析的通用框架对胶质瘤不同视角基因组数据进行分析。Wang等人[13]利用稀疏多元回归与稀疏典型相关分析之间的显式联系提出了基于特征向量的稀疏典型相关分析，研究甲状腺组织学图像和基因表达数据的相关性。

作为CCA的一种广义化扩展，多视角典型相关分析(Multi-view Canonical Correlation Analysis, MCCA)[14]能够对多个样本集之间的相关性进行表示，在不同研究中，多视角也通常被称为多模态或多重集等。MCCA中最佳线性变换可以通过求解广义特征值问题来获得，这对于高维数据来说计算量很大，样本的协方差矩阵也往往具有奇异性，这使得求解相关广义特征值问题具有挑战性。另外，MCCA只能以全局方式获得样本对之间的线性相关性，无法处理复杂的非线性情况。作为一种无监督的方法，MCCA没有利用监督信息，导致分类性能有限。

在过去的几十年里，MCCA为了解决这些限制已经扩展出了许多不同的新方法。当特征数量超过样本数量的情况时会导致协方差矩阵的奇异性，为了应对这一情况提出了正则化MCCA(Regularized Multi-view Canonical Correlation Analysis, RMCCA)[15]。利用正则化思想，通过图诱导嵌入多表示数据的几何结构信息构建了图正则化MCCA (Graph regularized Multiset Canonical Correlations, GrMCCs)[16]，在人脸数据集上的应用表明GrMCCs方法能够获得较佳实验结果。核MCCA (Kernel Multi-view Canonical Correlation Analysis, KMCCA)[17]是MCCA的一种流行的非线性扩展，将原始非线性数据隐式映射到高维特征空间，使其具有线性可分性，从而在高维空间中执行线性典型相关分析方法。利用样本类别信息构建视角间的鉴别相关性，提出了鉴别型MCCA(Discriminative Multi-view Canonical Correlation Analysis, DMCCA)[18]，使得子空间中具有紧密的类内分布以及类间离散分布，从而提高了低维特征的鉴别能力，在人类情感识别方面展现出良好性能。基于标签的MCCA(Labeled Multi-view Canonical Correlation Analysis, LMCCA)[19]充分利用训练样本的类内散布矩阵和多变量互相关矩阵来提取鉴别信

息，建立了基于类内信息进行典型相关分析的统一框架，该方法在人脸识别和利用多重特征目标识别等应用验证了其有效性。

目前已有的各种多视角学习方法主要是通过不同的优化准则将多视角数据投影到子空间，从而保留原始数据的有效鉴别特征，但是在利用子空间进行学习时，往往忽略了投影前后样本之间的相似性[20]，相似度顺序保持是一种重要的数据性质，它能够利用样本间的相似性来构建稳定的样本结构，因此本文提出一种相似度顺序保持特征提取方法，即相似度顺序保持跨视角相关分析(Similarity Order Preserving Across-view Correlation Analysis, SOPACA)，SOPACA能够利用高维训练样本为每个视角学习到一组投影方向，通过将原始高维基因样本投影到相似保序子空间，从而获得更具鉴别力的相似性特征，通过构建鉴别敏感的视角内相似度顺序保持散布和约束鉴别敏感的视角间相似度相关，使得相似性特征在保持投影前后样本两两之间的相似性的同时具有类内聚集性与类间离散性，不仅保持了样本之间的结构关系而且充分利用样本监督信息。本文在肺癌及结直肠癌基因表达数据上进行针对性实验，实验结果表明本方法的优越性。

2　多视角典型相关分析

给定经过中心化处理的m组高维基因数据集$\{\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)},\cdots,\boldsymbol{X}^{(m)}\}$，其中$\boldsymbol{X}^{(i)}=(\boldsymbol{x}_1^{(i)},\boldsymbol{x}_2^{(i)},\cdots,\boldsymbol{x}_n^{(i)})$ $\boldsymbol{R}^{d_i\times n}$，$n$为基因样本数，$d_i$为第$i$组特征的维数，$i=1,2,\cdots,m$, $\alpha^{(i)}\in R^{d_i}$表示与$\boldsymbol{X}^{(i)}$对应的相关投影方向。MCCA方法旨在为每个基因数据集寻找一组投影方向，使得投影后基因数据集间具有最大相关性，MCCA方法的优化准则为

$$\max_{\boldsymbol{\alpha}^{(1)},\boldsymbol{\alpha}^{(2)},\cdots,\boldsymbol{\alpha}^{(m)}}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_{ij}\boldsymbol{\alpha}^{(j)}}{\sqrt{\sum_{i=1}^{n}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_{ii}\boldsymbol{\alpha}^{(i)}}} \tag{1}$$

其中，$\boldsymbol{S}_{ij}$表示数据集$\boldsymbol{X}^{(i)}$和$\boldsymbol{X}^{(j)}$的视角间协方差矩阵，矩阵大小为$d_i\times d_j$，其计算形式为$\boldsymbol{S}_{ij}=\frac{1}{n}\sum_{u=1}^{n}(\boldsymbol{x}_u^{(i)}-\bar{x}^{(i)})(\boldsymbol{x}_u^{(j)}-\bar{x}^{(j)})^{\mathrm{T}}$，$\boldsymbol{S}_{ii}$为数据集$\boldsymbol{X}^{(i)}$视角内协方差矩阵，矩阵大小为$d_i\times d_i$，其计算形式为$\boldsymbol{S}_{ii}=\frac{1}{n}\sum_{u=1}^{n}(\boldsymbol{x}_u^{(i)}-\bar{x}^{(i)})(\boldsymbol{x}_u^{(i)}-\bar{x}^{(i)})^{\mathrm{T}}$, $\bar{x}^{(i)}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_u^{(i)}$和$\bar{x}^{(j)}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_u^{(j)}$分别对应$\boldsymbol{X}^{(i)}$和$\boldsymbol{X}^{(j)}$的均值。

由于投影方向具有尺度不变性，MCCA方法可以表述为如式(2)所示优化问题

$$\max_{\boldsymbol{\alpha}^{(1)},\boldsymbol{\alpha}^{(2)},\cdots,\boldsymbol{\alpha}^{(m)}}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_{ij}\boldsymbol{\alpha}^{(j)},$$
$$\text{s.t.}\ \sum_{i=1}^{n}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_{ii}\boldsymbol{\alpha}^{(i)}=1 \tag{2}$$

视角间协方差矩阵$\boldsymbol{S}_{ij}$反映了数据集$\boldsymbol{X}^{(i)}$和$\boldsymbol{X}^{(j)}$之间的相关性，视角内协方差矩阵$\boldsymbol{S}_{ii}$反映了数据集$\boldsymbol{X}^{(i)}$的整体散布信息，MCCA方法可以视为在最大化视角间相关性的同时最小化视角内散布信息。

3　相似度顺序保持跨视角相关分析

3.1　构建鉴别敏感的视角内相似度顺序保持散布

假设第i组基因样本$\boldsymbol{x}_1^{(i)},\boldsymbol{x}_2^{(i)},\cdots,\boldsymbol{x}_n^{(i)}$在相关投影方向$\boldsymbol{\alpha}^{(i)}$上的相关特征为$\boldsymbol{y}_1^{(i)},\boldsymbol{y}_2^{(i)},\cdots,\boldsymbol{y}_n^{(i)}$，即$\boldsymbol{y}_l^{(i)}=\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{x}_l^{(i)}$，其中$l=1,2,\cdots,n$且$i=1,2,\cdots,m$。

为了在相似保序子空间中保持同类基因样本的相似性顺序并且投影后同类基因样本之间尽可能聚集，目标函数可以表示为

$$\min_{(i)}\sum_{i=1}^{m}\sum_{u=1}^{n}\sum_{c(\boldsymbol{x}_v^{(i)})=c(\boldsymbol{x}_u^{(i)})}\sum_{c(\boldsymbol{x}_k^{(i)})=c(\boldsymbol{x}_u^{(i)})}\left(\boldsymbol{q}_{uv}^{(i)}-\boldsymbol{q}_{uk}^{(i)}\right)$$
$$\times\left(\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)}\right\|_2^2-\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_k^{(i)}\right\|_2^2\right) \tag{3}$$

其中，$c(\boldsymbol{x}_v^{(i)})=c(\boldsymbol{x}_u^{(i)})$表示样本$\boldsymbol{x}_v^{(i)}$与$\boldsymbol{x}_u^{(i)}$属于同一类，$\boldsymbol{q}_{uv}^{(i)}$表示第$i$组基因数据集中任意两个样本$\boldsymbol{x}_u^{(i)}$和$\boldsymbol{x}_v^{(i)}$之间的归一化相似度，具体表示为

$$\boldsymbol{q}_{uv}^{(i)}=\mathrm{e}^{-\sigma\|\boldsymbol{x}_u^{(i)}-\boldsymbol{x}_v^{(i)}\|_2^2} \tag{4}$$

度量值$\boldsymbol{q}_{uv}^{(i)}$越大，表明样本之间具有较大相似度。$\boldsymbol{q}_{uv}^{(i)}-\boldsymbol{q}_{uk}^{(i)}$表示同一视角下同类样本$\boldsymbol{x}_u^{(i)},\boldsymbol{x}_v^{(i)}$与$\boldsymbol{x}_u^{(i)},\boldsymbol{x}_k^{(i)}$之间的相似度差值，最小化$\boldsymbol{q}_{uv}^{(i)}-\boldsymbol{q}_{uk}^{(i)}$达到保持同类样本的相似性顺序的目的。

由于$\boldsymbol{q}_{uv}^{(i)}$只能度量两个样本之间的相似度，因此定义相似度矩阵对两两样本点之间的相似度进行计算，对于任意两个同类样本点$\boldsymbol{x}_u^{(i)}$和$\boldsymbol{x}_v^{(i)}$，寻找与它们属于同一类的所有样本点，首先计算$\boldsymbol{x}_u^{(i)}$和$\boldsymbol{x}_v^{(i)}$之间样本相似度，再计算$\boldsymbol{x}_u^{(i)}$与剩余同类样本点的样本相似度，将二者计算所得结果之差累加得到矩阵中的一个元素，构建相似度矩阵$\boldsymbol{Q}_{uv}^{(i)}$

$$\boldsymbol{Q}_{uv}^{(i)}=\begin{cases}\sum_{c(\boldsymbol{x}_u^{(i)})=c(\boldsymbol{x}_k^{(i)})}\left(\boldsymbol{q}_{uv}^{(i)}-\boldsymbol{q}_{uk}^{(i)}\right), c(\boldsymbol{x}_v^{(i)})=c(\boldsymbol{x}_u^{(i)})\\ 0, \qquad\qquad\qquad\qquad \text{其他}\end{cases} \tag{5}$$

利用相似度矩阵对式(3)进行化简，经过代数变换，式(3)能够推导为

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{u=1}^{n}\sum_{c(\boldsymbol{x}_v^{(i)})=c(\boldsymbol{x}_u^{(i)})}\sum_{c(\boldsymbol{x}_k^{(i)})=c(\boldsymbol{x}_u^{(i)})}\left(\boldsymbol{q}_{uv}^{(i)}-\boldsymbol{q}_{uk}^{(i)}\right)\times\left(\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)}\right\|_2^2-\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_k^{(i)}\right\|_2^2\right)\\
&=\sum_{i=1}^{m}\sum_{u=1}^{n}\sum_{c(\boldsymbol{x}_v^{(i)})=c(\boldsymbol{x}_u^{(i)})}\sum_{c(\boldsymbol{x}_k^{(i)})=c(\boldsymbol{x}_u^{(i)})}\left(\left(\boldsymbol{q}_{uv}^{(i)}-\boldsymbol{q}_{uk}^{(i)}\right)\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)}\right\|_2^2-\left(\boldsymbol{q}_{uv}^{(i)}-\boldsymbol{q}_{uk}^{(i)}\right)\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)}\right\|_2^2\right)\\
&=\sum_{i=1}^{m}\sum_{u=1}^{n}\sum_{c(\boldsymbol{x}_v^{(i)})=c(\boldsymbol{x}_u^{(i)})}\sum_{c(\boldsymbol{x}_k^{(i)})=c(\boldsymbol{x}_u^{(i)})}\left(\left(\boldsymbol{q}_{uv}^{(i)}-\boldsymbol{q}_{uk}^{(i)}\right)\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)}\right\|_2^2+\left(\boldsymbol{q}_{uk}^{(i)}-\boldsymbol{q}_{uv}^{(i)}\right)\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)}\right\|_2^2\right)\\
&=\sum_{i=1}^{m}\sum_{u=1}^{n}\sum_{c(\boldsymbol{x}_v^{(i)})=c(\boldsymbol{x}_u^{(i)})}\boldsymbol{Q}_{uv}^{(i)}\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)}\right\|_2^2+\sum_{i=1}^{m}\sum_{u=1}^{n}\sum_{c(\boldsymbol{x}_k^{(i)})=c(\boldsymbol{x}_u^{(i)})}\boldsymbol{Q}_{uk}^{(i)}\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)}\right\|_2^2\\
&=\sum_{i=1}^{m}\sum_{u=1}^{n}\sum_{c(\boldsymbol{x}_v^{(i)})=c(\boldsymbol{x}_u^{(i)})}\boldsymbol{Q}_{uv}^{(i)}\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)}\right\|_2^2+\sum_{i=1}^{m}\sum_{u=1}^{n}\sum_{c(\boldsymbol{x}_v^{(i)})=c(\boldsymbol{x}_u^{(i)})}\boldsymbol{Q}_{uv}^{(i)}\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)}\right\|_2^2\\
&=2\sum_{i=1}^{m}\sum_{u=1}^{n}\sum_{v=1}^{n}\boldsymbol{Q}_{uv}^{(i)}\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)}\right\|_2^2
\end{aligned}\tag{6}$$

由于式(6)结果无法直接优化求解，因此将其进一步推导为矩阵形式

$$\begin{aligned}
&2\sum_{i=1}^{m}\sum_{u=1}^{n}\sum_{v=1}^{n}\boldsymbol{Q}_{uv}^{(i)}\left\|\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)}\right\|_2^2\\
&=2\sum_{i=1}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\left(\sum_{u=1}^{n}\sum_{v=1}^{n}\boldsymbol{Q}_{uv}^{(i)}\left(\boldsymbol{x}_u^{(i)}-\boldsymbol{x}_v^{(i)}\right)\left(\boldsymbol{x}_u^{(i)}-\boldsymbol{x}_v^{(i)}\right)^{\mathrm{T}}\right)\boldsymbol{\alpha}^{(i)}\\
&=2\sum_{i=1}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\sum_{u=1}^{n}\sum_{v=1}^{n}\boldsymbol{Q}_{uv}^{(i)}\left(\boldsymbol{x}_u^{(i)}\boldsymbol{x}_u^{(i)\mathrm{T}}-\boldsymbol{x}_u^{(i)}\boldsymbol{x}_v^{(i)\mathrm{T}}-\boldsymbol{x}_v^{(i)}\boldsymbol{x}_u^{(i)\mathrm{T}}+\boldsymbol{x}_v^{(i)}\boldsymbol{x}_v^{(i)\mathrm{T}}\right)\boldsymbol{\alpha}^{(i)}\\
&=2\sum_{i=1}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{X}^{(i)}\left(2\boldsymbol{D}^{(i)}-\boldsymbol{Q}^{(i)}-\boldsymbol{Q}^{(i)\mathrm{T}}\right)\boldsymbol{X}^{(i)\mathrm{T}}\boldsymbol{\alpha}^{(i)}\\
&=2\sum_{i=1}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{X}^{(i)}\boldsymbol{L}_w^{(i)}\boldsymbol{X}^{(i)\mathrm{T}}\boldsymbol{\alpha}^{(i)}\\
&=2\sum_{i=1}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_w^{(i)}\boldsymbol{\alpha}^{(i)}
\end{aligned}\tag{7}$$

其中，$\boldsymbol{D}^{(i)}$为对角矩阵，对角元素表示为$\boldsymbol{D}_{uu}^{(i)}=\frac{1}{2}\sum_{v=1}^{n}(\boldsymbol{Q}_{uv}^{(i)}+\boldsymbol{Q}_{vu}^{(i)})$,$\boldsymbol{L}_w^{(i)}=2\boldsymbol{D}^{(i)}-\boldsymbol{Q}^{(i)}-\boldsymbol{Q}^{(i)\mathrm{T}}$为拉普拉斯矩阵，$\boldsymbol{S}_w^{(i)}=\boldsymbol{X}^{(i)}\boldsymbol{L}_w^{(i)}\boldsymbol{X}^{(i)\mathrm{T}}$为视角内相似度顺序保持散布矩阵，通过构建鉴别敏感的视角内相似度顺序保持散布，可以保持视角类内紧凑性，由于常数对于投影方向$\boldsymbol{\alpha}^{(i)}$求解没有影响，因此可以被省略，目标函数可以表示为$\min\limits_{\boldsymbol{\alpha}^{(i)}}\sum_{i=1}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_w^{(i)}\cdot\boldsymbol{\alpha}^{(i)}$。

3.2 构建鉴别敏感的视角间相似度相关

首先将MCCA方法中协方差矩阵$\boldsymbol{S}_{ij}$推导为如式(8)的等价形式[16]

$$\boldsymbol{S}_{ij}=\frac{1}{2n^2}\sum_{u=1}^{n}\sum_{v=1}^{n}\left(\boldsymbol{x}_u^{(i)}-\boldsymbol{x}_v^{(i)}\right)\left(\boldsymbol{x}_u^{(j)}-\boldsymbol{x}_v^{(j)}\right)^{\mathrm{T}}\tag{8}$$

则式(1)中视角间相关性可以表述为

$$\rho_{ij}\left(\boldsymbol{\alpha}^{(i)},\boldsymbol{\alpha}^{(j)}\right)=\frac{1}{2n^2}\sum_{u=1}^{n}\sum_{v=1}^{n}\boldsymbol{\alpha}^{(i)\mathrm{T}}\left(\boldsymbol{x}_u^{(i)}-\boldsymbol{x}_v^{(i)}\right)\cdot\left(\boldsymbol{x}_u^{(j)}-\boldsymbol{x}_v^{(j)}\right)^{\mathrm{T}}\boldsymbol{\alpha}^{(j)}\tag{9}$$

其中，$\rho_{ij}(\boldsymbol{\alpha}^{(i)},\boldsymbol{\alpha}^{(j)})=\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_{ij}\boldsymbol{\alpha}^{(j)}$表示相关特征$\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{X}^{(i)}$和$\boldsymbol{\alpha}^{(j)\mathrm{T}}\boldsymbol{X}^{(j)}(i\neq j)$之间的相关性。式(9)中视角间相关性是基于视角内成对数据的等价表示，该形式更利于视角间相似度相关矩阵的构建。由于常数对于投影方向的求解没有影响，MCCA的目标函数可以表示为

$$\max_{\boldsymbol{\alpha}^{(1)},\boldsymbol{\alpha}^{(2)},\cdots,\boldsymbol{\alpha}^{(m)}}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\alpha}^{(i)\mathrm{T}}\left(\sum_{u=1}^{n}\sum_{v=1}^{n}\left(\boldsymbol{x}_u^{(i)}-\boldsymbol{x}_v^{(i)}\right)\cdot\left(\boldsymbol{x}_u^{(j)}-\boldsymbol{x}_v^{(j)}\right)^{\mathrm{T}}\right)\boldsymbol{\alpha}^{(j)}\tag{10}$$

对于任意两组基因视角数据$\boldsymbol{X}^{(i)}=(\boldsymbol{x}_1^{(i)},\boldsymbol{x}_2^{(i)},\cdots,$

$\boldsymbol{x}_n^{(i)}$)与$\boldsymbol{X}^{(j)}=(\boldsymbol{x}_1^{(j)},\boldsymbol{x}_2^{(j)},\cdots,\boldsymbol{x}_n^{(j)})(i\neq j)$，利用基因样本的相似性来构造类间相似性加权矩阵，其定义为

$$\boldsymbol{W}_{uv}^{(i)}=\begin{cases}\boldsymbol{Q}_{uv}^{(i)}, c(\boldsymbol{x}_u^{(i)})\neq c(\boldsymbol{x}_v^{(i)})\\ 0, \quad \text{其他}\end{cases} \tag{11}$$

$$\begin{aligned}&\max_{\boldsymbol{\alpha}^{(i)},\boldsymbol{\alpha}^{(j)}}\sum_{u=1}^{n}\sum_{v=1}^{n}\left[\boldsymbol{W}_{uv}^{(i)}(\boldsymbol{y}_u^{(i)}-\boldsymbol{y}_v^{(i)})\right]\left[\boldsymbol{W}_{uv}^{(j)}(\boldsymbol{y}_u^{(j)}-\boldsymbol{y}_v^{(j)})^{\mathrm{T}}\right]\\ &=\boldsymbol{\alpha}^{(i)\mathrm{T}}\left(\sum_{u=1}^{n}\sum_{v=1}^{n}\boldsymbol{W}_{uv}^{(i)}\boldsymbol{W}_{uv}^{(j)}(\boldsymbol{x}_u^{(i)}-\boldsymbol{x}_v^{(i)})(\boldsymbol{x}_u^{(j)}-\boldsymbol{x}_v^{(j)})^{\mathrm{T}}\right)\boldsymbol{\alpha}^{(j)}\\ &=\boldsymbol{\alpha}^{(i)\mathrm{T}}\left(\sum_{u=1}^{n}\sum_{v=1}^{n}\boldsymbol{W}_{uv}^{(i)}\boldsymbol{W}_{uv}^{(j)}(\boldsymbol{x}_u^{(i)}\boldsymbol{x}_u^{(j)\mathrm{T}}-\boldsymbol{x}_u^{(i)}\boldsymbol{x}_v^{(j)\mathrm{T}}-\boldsymbol{x}_v^{(i)}\boldsymbol{x}_u^{(j)\mathrm{T}}+\boldsymbol{x}_v^{(i)}\boldsymbol{x}_v^{(j)\mathrm{T}})\right)\boldsymbol{\alpha}^{(j)}\\ &=\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{X}^{(i)}(2\boldsymbol{D}^{(ij)}-\boldsymbol{W}^{(ij)}-\boldsymbol{W}^{(ij)\mathrm{T}})\boldsymbol{X}^{(j)\mathrm{T}}\boldsymbol{\alpha}^{(j)}\\ &=\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{X}^{(i)}\boldsymbol{L}_b^{(ij)}\boldsymbol{X}^{(j)\mathrm{T}}\boldsymbol{\alpha}^{(j)}\\ &=\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_b^{(ij)}\boldsymbol{\alpha}^{(j)}\end{aligned} \tag{13}$$

其中，$\boldsymbol{D}^{(ij)}$为对角矩阵，对角元素为$\boldsymbol{D}_{uu}^{(ij)}=\frac{1}{2}\sum_{v=1}^{n}(\boldsymbol{W}_{uv}^{(i)}\boldsymbol{W}_{uv}^{(j)}+\boldsymbol{W}_{vu}^{(i)}\boldsymbol{W}_{vu}^{(j)})$，$\boldsymbol{L}_b^{(ij)}=2\boldsymbol{D}^{(ij)}-\boldsymbol{W}^{(ij)}-\boldsymbol{W}^{(ij)\mathrm{T}}$为拉普拉斯矩阵，$\boldsymbol{W}^{(ij)}=\boldsymbol{W}^{(i)}\circ\boldsymbol{W}^{(j)}$，运算$\circ$表示矩阵对应元素相乘，$\boldsymbol{S}_b^{(ij)}=\boldsymbol{X}^{(i)}\boldsymbol{L}_b^{(ij)}\boldsymbol{X}^{(j)\mathrm{T}}$表示视角间相似度相关矩阵。

3.3 SOPACA的建模与求解

SOPACA方法期望学习的相关特征能够在保持不同视角间特征类内紧凑性和相似度顺序的同时具有较大的类间距离，这种期望可以表述为

$$\left.\begin{aligned}&\max_{\boldsymbol{\alpha}^{(i)},\boldsymbol{\alpha}^{(j)}}\sum_{i=1}^{m}\sum_{j=1,i\neq j}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_b^{(ij)}\boldsymbol{\alpha}^{(j)}\\ &\min_{\boldsymbol{\alpha}^{(i)}}\sum_{i=1}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_w^{(i)}\boldsymbol{\alpha}^{(i)}\end{aligned}\right\} \tag{14}$$

通过最大化视角间相似度相关并且最小化视角内相似度顺序保持散布，将其与多视角典型相关分析的目标函数相融合，从而构建出相似度顺序保持跨视角相关分析方法，借助常用优化模型构造方法[21,22]，SOPACA的优化问题可以描述为

$$\begin{aligned}&\max J(\boldsymbol{\alpha})=\sum_{i=1}^{m}\sum_{j=1,i\neq j}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_b^{(ij)}\boldsymbol{\alpha}^{(j)},\\ &\text{s.t.}\sum_{i=1}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_w^{(i)}\boldsymbol{\alpha}^{(i)}=1\end{aligned} \tag{15}$$

其中，$\boldsymbol{\alpha}^{(i)\mathrm{T}}=\boldsymbol{\alpha}^{(1)\mathrm{T}},\boldsymbol{\alpha}^{(2)\mathrm{T}},\cdots,\boldsymbol{\alpha}^{(m)\mathrm{T}}$。

为了对式(15)进行优化求解，构建如式(16)所示Lagrange函数

$$\begin{aligned}F(\boldsymbol{\alpha},\lambda)=&\sum_{i=1}^{m}\sum_{j=1,i\neq j}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_b^{(ij)}\boldsymbol{\alpha}^{(j)}\\ &-\lambda\left(\sum_{l=1}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_w^{(i)}\boldsymbol{\alpha}^{(i)}-1\right)\end{aligned} \tag{16}$$

$$\boldsymbol{W}_{uv}^{(j)}=\begin{cases}\boldsymbol{Q}_{uv}^{(j)}, c(\boldsymbol{x}_u^{(j)})\neq c(\boldsymbol{x}_v^{(j)})\\ 0, \quad \text{其他}\end{cases} \tag{12}$$

为了使得投影后相似保序子空间中不同类别的基因样本之间具有较大的类间距离，目标函数表示如式(13)所示

其中，λ为Lagrange乘子，令$\partial F/\partial\boldsymbol{\alpha}^{(i)}=0$，有

$$\frac{\partial F}{\partial\boldsymbol{\alpha}^{(i)}}=\sum_{j=1}^{m}\boldsymbol{S}_b^{(ij)}\boldsymbol{\alpha}^{(j)}-\lambda\boldsymbol{S}_w^{(i)}\boldsymbol{\alpha}^{(i)}=0, i=1,2,\cdots,m \tag{17}$$

式(17)左乘$\boldsymbol{\alpha}^{(i)\mathrm{T}}(i=1,2,\cdots,m)$，可得

$$\sum_{i=1}^{m}\sum_{\substack{j=1\\ i\neq j}}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_b^{(ij)}\boldsymbol{\alpha}^{(j)}-\lambda\sum_{i=1}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_w^{(i)}\boldsymbol{\alpha}^{(i)}=0 \tag{18}$$

进而可得

$$\lambda=\frac{\sum_{i=1}^{m}\sum_{j=1,i\neq j}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_b^{(ij)}\boldsymbol{\alpha}^{(j)}}{\sum_{i=1}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_w^{(i)}\boldsymbol{\alpha}^{(i)}} \tag{19}$$

根据$\sum_{i=1}^{m}\boldsymbol{\alpha}^{(i)\mathrm{T}}\boldsymbol{S}_w^{(i)}\boldsymbol{\alpha}^{(i)}=1$及式(19)可知，$\lambda$即可表示SOPACA的目标函数，将式(17)转化为式(20)所示的广义特征值问题

$$\begin{aligned}&\begin{bmatrix}&\boldsymbol{S}_b^{(12)}&\cdots&\boldsymbol{S}_b^{(1m)}\\ \boldsymbol{S}_b^{(21)}&&\cdots&\boldsymbol{S}_b^{(2m)}\\ \vdots&\vdots&\ddots&\vdots\\ \boldsymbol{S}_b^{(m1)}&\boldsymbol{S}_b^{(m2)}&\cdots&\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}^{(1)}\\ \boldsymbol{\alpha}^{(2)}\\ \vdots\\ \boldsymbol{\alpha}^{(m)}\end{bmatrix}\\ &=\lambda\begin{bmatrix}\boldsymbol{S}_w^{(1)}&&&\\ &\boldsymbol{S}_w^{(2)}&&\\ &&\ddots&\\ &&&\boldsymbol{S}_w^{(m)}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}^{(1)}\\ \boldsymbol{\alpha}^{(2)}\\ \vdots\\ \boldsymbol{\alpha}^{(m)}\end{bmatrix}\end{aligned} \tag{20}$$

通过对式(20)进行求解，能够获得前d个最大特征值对应的特征向量$\{\boldsymbol{\alpha}_k^{\mathrm{T}}=(\boldsymbol{\alpha}_k^{(1)\mathrm{T}},\boldsymbol{\alpha}_k^{(2)\mathrm{T}},\cdots,\boldsymbol{\alpha}_k^{(m)\mathrm{T}})\}_{k=1}^{d}$作为解向量，最终得到对应$m$组数据的$m$个投影矩阵$\{\boldsymbol{W}_i=(\boldsymbol{\alpha}_1^{(i)},\boldsymbol{\alpha}_2^{(i)},\cdots,\boldsymbol{\alpha}_d^{(i)})\}_{i=1}^{m}$，分别

利用$\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{X}_1,\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{X}_2,\cdots,\boldsymbol{W}_m^{\mathrm{T}}\boldsymbol{X}_m$抽取特征，采用式(21)所示并行融合策略进行特征融合以形成有效的鉴别矢量$\boldsymbol{Z}$

$$\begin{aligned}\boldsymbol{Z}&=\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{X}_1+\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{X}_2+\cdots+\boldsymbol{W}_m^{\mathrm{T}}\boldsymbol{X}_m\\&=\begin{bmatrix}\boldsymbol{W}_1\\\boldsymbol{W}_2\\\vdots\\\boldsymbol{W}_m\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{X}_1\\\boldsymbol{X}_2\\\vdots\\\boldsymbol{X}_m\end{bmatrix}\end{aligned}\tag{21}$$

式(21)形成的鉴别矢量$\boldsymbol{Z}$将代表m组高维基因数据应用于分类任务，使用基于欧氏距离的最近邻分类器[23]进行分类识别，本文所提SOPACA方法步骤如算法1所示。

4 实验结果与分析

为了验证SOPACA方法在癌症分类上的有效性，分别在肺癌和结直肠癌基因表达数据集上进行实验来评估SOPACA方法的识别性能。使用模态策略[24]获得基因表达数据的3种模态数据，具体而言，将基因表达数据看作时序信号，分别使用Coiflets,Daubechies和Symlets 3种小波变换提取其低频分量作为3种模态数据，由于基因表达数据具有高维与小样本之间的不平衡问题，使用PCA方法对模态数据维数统一约减至100维以保证实验的稳定性。实验中将SOPACA方法与LMCCA[19], GrMCCs[16], MCCA[14], LDA[5]方法进行对比分析，采用基于欧氏距离的最近邻分类器对基因样本进行分类得到最终识别结果。

算法1 SOPACA方法步骤

输入：视角数据集$\{\boldsymbol{X}^{(i)}=(\boldsymbol{x}_1^{(i)},\boldsymbol{x}_2^{(i)},\cdots,\boldsymbol{x}_n^{(i)})\in\boldsymbol{R}^{d_i\times n}\}_{i=1}^m$

输出：基因样本类标签

(1)利用式(7)和式(13)分别构建视角内相似度顺序保持散布矩阵$\boldsymbol{S}_w^{(i)}$和视角间相似度相关矩阵$\boldsymbol{S}_b^{(ij)}$；

(2)利用式(16)Lagrange函数求得特征值λ和对应特征向量；

(3)利用式(20)获得相关投影矩阵$\{\boldsymbol{W}_i=(\boldsymbol{\alpha}_1^{(i)},\boldsymbol{\alpha}_2^{(i)},\cdots,\boldsymbol{\alpha}_d^{(i)})\}_{i=1}^m$；

(4)利用式(21)获得特征融合后的鉴别矢量$\boldsymbol{Z}$；

(5)利用基于欧氏距离的最近邻分类器对鉴别矢量$\boldsymbol{Z}$进行分类，得到基因样本类标签。

4.1 在肺癌基因表达数据集上的实验

肺癌基因表达数据集包含107个样本，每个样本均包含22283个探针测得的基因表达水平(下载地址为：https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10072)，其中癌症样本为58个，正常样本为49个。在实验中，随机抽取每类t(t=5, 10, 15, 20, 25)个样本构建训练集，其余样本作为测试集，每个实验独立运行10次，最终得到如表1所示各算法在肺癌基因表达数据集上的平均识别率以及对应识别率的标准差。

MCCA只能全局地获取样本之间的线性相关性，因此在复杂的非线性情景下往往对数据拟合不足导致识别性能有限。LDA作为有监督的单视图学习方法，识别效果略优于MCCA方法。GrMCCs引入图结构考虑了多视角数据的几何结构，由于原始基因表达数据包含大量冗余信息及噪声，导致数据真实局部几何结构失真从而影响低维特征的鉴别力，这种失真关系使得GrMCCs的部分识别率低于MCCA。LMCCA充分利用样本类别信息获得了较高识别率，SOPACA方法保持了投影前后样本结构关系，充分利用类别信息获得类内紧凑性与类间分散性，随着样本数量增加SOPACA方法始终保持了最优识别率，在样本数量较小的情况下，SOPACA方法相较于其他算法显示出其优越性，对于具有高维小样本特点的基因表达数据尤为重要。标准差能够反映识别率的波动情况，标准差越大说明数据波动性越强，与其他方法相比，SOPACA方法拥有较小标准差表明识别率变化平缓，说明所提出的方法具有良好的鲁棒性。

4.2 在结直肠癌基因表达数据集上的实验

结直肠癌基因表达数据集包含34个样本，每个样本均包含54675个探针测得的基因表达水平(下载地址为：https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32323)，其中癌症样本和正常样本均为17个。在实验中随机抽取每类t(t=2, 3, 4, 5, 6)个样本作为训练样本，其余样本用于测试，每个实验独立重复10次得到如表2所示各算法在结直肠癌数据集上的平均识别率以及对应识别率标准差。

表 1 在肺癌基因表达数据集上的识别率变化结果

方法	5训练样本	10训练样本	15训练样本	20训练样本	25训练样本
SOPACA	**98.66±0.85**	**99.08±0.91**	**98.70±1.22**	**98.81±0.94**	**99.65±0.74**
MCCA	96.08±2.37	98.16±1.11	97.92±1.40	97.61±1.00	99.30±1.11
LDA	96.70±2.05	98.05±1.22	98.31±1.23	98.51±1.00	99.30±0.91
GrMCCs	94.64±3.10	96.55±2.82	98.05±2.31	97.61±2.01	98.60±1.38
LMCCA	97.01±1.41	98.28±1.12	98.18±1.40	98.36±1.10	99.30±0.91

表 2　在结直肠癌基因表达数据集上的平均识别率

方法	2训练样本	3训练样本	4训练样本	5训练样本	6训练样本
SOPACA	**98.67±1.72**	**99.29±1.51**	**99.23±1.62**	**99.58±1.32**	**99.09±1.92**
MCCA	95.67±3.52	97.50±2.41	97.31±2.60	99.17±1.76	98.18±2.35
LDA	95.00±2.83	96.07±2.03	95.77±4.23	96.67±2.64	97.73±2.40
GrMCCs	93.33±8.75	94.29±2.50	96.92±3.53	97.50±2.15	98.64±2.20
LMCCA	96.67±2.22	96.07±2.41	97.31±3.17	97.50±2.15	97.73±2.40

MCCA方法只能处理简单的线性问题，无法提取更具鉴别力的特征。LDA表现出了较低识别率结果，反映了多视角学习方法对于特征抽取的优越性。与实验4.1中结果类似，利用图正则化技术的GrMCCs方法由于局部失真带来的影响导致平均识别率较低。DMCCA方法利用样本类别信息构建视角间的鉴别相关性，LMCCA充分利用训练样本的类内散布矩阵和多变量互相关矩阵来提取鉴别信息，二者都取得了较好的识别性能。SOPACA方法通过将样本投影到相似保序子空间能够获得更具鉴别力的特征，不仅保持了样本之间的结构关系而且充分利用样本类别信息，实验结果表明SOPACA方法的识别精度始终优于其他对比算法，当训练样本数量较少时更能体现出相似度顺序保持的优势。

5　结束语

传统基于子空间投影的多视角学习方法往往会忽略投影前后样本之间的相似性，进而影响多视角学习性能。本文提出了SOPACA方法，通过将基因表达数据投影到相似保序子空间，该子空间中的低维数据能够在保持投影前后样本相似度的情况下具有类内聚集性与类间离散性，从而有效增强低维数据的鉴别能力，在维持样本结构关系的同时充分利用了样本监督信息。在基因表达数据集上的实验表明，本文算法抽取的特征相较于其他特征提取算法更具鉴别性。

参 考 文 献

[1] SHUMATE A and SALZBERG S L. Liftoff: Accurate mapping of gene annotations[J]. *Bioinformatics*, 2021, 37(12): 1639–1643. doi: 10.1093/BIOINFORMATICS/BTAA1016.

[2] LU Rongxiu, CAI Yingjie, ZHU Jianyong, *et al*. Dimension reduction of multimodal data by auto-weighted local discriminant analysis[J]. *Neurocomputing*, 2021, 461: 27–40. doi: 10.1016/J.NEUCOM.2021.06.035.

[3] 王肖锋, 孙明月, 葛为民. 基于图像协方差无关的增量特征提取方法研究[J]. 电子与信息学报, 2019, 41(11): 2768–2776. doi: 10.11999/JEIT181138.
WANG Xiaofeng, SUN Mingyue, and GE Weimin. An incremental feature extraction method without estimating image covariance matrix[J]. *Journal of Electronics & Information Technology*, 2019, 41(11): 2768–2776. doi: 10.11999/JEIT181138.

[4] ARTONI F, DELORME A, and MAKEIG S. Applying dimension reduction to EEG data by principal component analysis reduces the quality of its subsequent independent component decomposition[J]. *NeuroImage*, 2018, 175: 176–187. doi: 10.1016/j.neuroimage.2018.03.016.

[5] LI Chunna, SHAO Yuanhai, CHEN Weijie, *et al*. Generalized two-dimensional linear discriminant analysis with regularization[J]. *Neural Networks*, 2021, 142: 73–91. doi: 10.1016/J.NEUNET.2021.04.030.

[6] NAKAYAMA Y, YATA K, and AOSHIMA M. Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings[J]. *Journal of Multivariate Analysis*, 2021, 185: 104779. doi: 10.1016/J.JMVA.2021.104779.

[7] CLAYMAN C L, SRINIVASAN S M, and SANGWAN R S. K-means clustering and principal components analysis of microarray data of L1000 landmark genes[J]. *Procedia Computer Science*, 2020, 168: 97–104. doi: 10.1016/j.procs.2020.02.265.

[8] WANG Cheng, CAO Longbing, and MIAO Baiqi. Optimal feature selection for sparse linear discriminant analysis and its applications in gene expression data[J]. *Computational Statistics & Data Analysis*, 2013, 66: 140–149. doi: 10.1016/j.csda.2013.04.003.

[9] LIN Weiming, GAO Qinquan, DU Min, *et al*. Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data[J]. *Computers in Biology and Medicine*, 2021, 134: 104478. doi: 10.1016/J.COMPBIOMED.2021.104478.

[10] 苏树智, 谢军, 平昕瑞, 等. 图强化典型相关分析及在图像识别中的应用[J]. 电子与信息学报, 2021, 43(11): 3342–3349. doi: 10.11999/JEIT210154.
SU Shuzhi, XIE Jun, PING Xinrui, *et al*. Graph enhanced canonical correlation analysis and its application to image recognition[J]. *Journal of Electronics & Information Technology*, 2021, 43(11): 3342–3349. doi: 10.11999/JEIT210154.

[11] LIN Dongdong, CALHOUN V D, and WANG Yuping. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis[J]. *Medical Image Analysis*, 2014, 18(6): 891–902. doi: 10.1016/j.media.2013.10.010.

[12] TENENHAUS A, PHILIPPE C, and FROUIN V. Kernel generalized canonical correlation analysis[J]. *Computational Statistics & Data Analysis*, 2015, 90: 114–131. doi: 10.1016/j.csda.2015.04.004.

[13] WANG Wenjia and ZHOU Yihui. Eigenvector-based sparse canonical correlation analysis: Fast computation for estimation of multiple canonical vectors[J]. *Journal of Multivariate Analysis*, 2021, 185: 104781. doi: 10.1016/J.JMVA.2021.104781.

[14] YUAN Yunhao, SUN Quansen, ZHOU Qiang, *et al*. A novel multiset integrated canonical correlation analysis framework and its application in feature fusion[J]. *Pattern Recognition*, 2011, 44(5): 1031–1040. doi: 10.1016/j.patcog.2010.11.004.

[15] DELEUS F and VAN HULLE M M. Functional connectivity analysis of fMRI data based on regularized multiset canonical correlation analysis[J]. *Journal of Neuroscience Methods*, 2011, 197(1): 143–157. doi: 10.1016/j.jneumeth.2010.11.029.

[16] YUAN Yunhao and SUN Quansen. Graph regularized multiset canonical correlations with applications to joint feature extraction[J]. *Pattern Recognition*, 2014, 47(12): 3907–3919. doi: 10.1016/j.patcog.2014.06.016.

[17] SU Shuzhi, GE Hongwei, and YUAN Yunhao. Kernel-aligned multi-view canonical correlation analysis for image recognition[J]. *Infrared Physics & Technology*, 2016, 78: 233–240. doi: 10.1016/j.infrared.2016.08.010.

[18] GAO Lei, QI Lin, CHEN Enqing, *et al*. Discriminative multiple canonical correlation analysis for information fusion[J]. *IEEE Transactions on Image Processing*, 2018, 27(4): 1951–1965. doi: 10.1109/TIP.2017.2765820.

[19] GAO Lei, ZHANG Rui, QI Lin, *et al*. The labeled multiple canonical correlation analysis for information fusion[J]. *IEEE Transactions on Multimedia*, 2019, 21(2): 375–387. doi: 10.1109/TMM.2018.2859590.

[20] HU Haoshuang, FENG Dazheng, and CHEN Qingyan. A novel dimensionality reduction method: Similarity order preserving discriminant analysis[J]. *Signal Processing*, 2021, 182: 107933. doi: 10.1016/J.SIGPRO.2020.107933.

[21] SU Shuzhi, ZHU Gang, and ZHU Yanmin. An orthogonal locality and globality dimensionality reduction method based on Twin Eigen decomposition[J]. *IEEE Access*, 2021, 9: 55714–55725. doi: 10.1109/ACCESS.2021.3071192.

[22] SHEN Xiaobo, SUN Quansen, and YUAN Yunhao. A unified multiset canonical correlation analysis framework based on graph embedding for multiple feature extraction[J]. *Neurocomputing*, 2015, 148: 397–408. doi: 10.1016/j.neucom.2014.06.015.

[23] SHOKRZADE A, RAMEZANI M, TAB F A, *et al*. A novel extreme learning machine based kNN classification method for dealing with big data[J]. *Expert Systems with Applications*, 2021, 183: 115293. doi: 10.1016/J.ESWA.2021.115293.

[24] LIU Dongwei, JIA Runping, WANG Caifeng, *et al*. Automated detection of cancerous genomic sequences using genomic signal processing and machine learning[J]. *Future Generation Computer Systems*, 2019, 98: 233–237. doi: 10.1016/J.FUTURE.2018.12.041.

苏树智：男，副教授，研究方向为多模态模式识别、特征学习、基因分析.

张开宇：男，硕士生，研究方向为多模态模式识别、基因分析.

王子莹：女，硕士生，研究方向为模式识别、图像处理.

张茂岩：男，硕士生，研究方向为模式识别.

责任编辑：余 蓉