基于稀疏类别保留投影的基因表达数据降维方法

王文俊
(西安电子科技大学计算机学院,陕西西安 710071)

摘 要: 针对基因表达数据高维小样本特性所带来的维数灾难问题,结合回归和类别保留投影方法,提出一种新的基因表达数据降维方法,叫稀疏类别保留投影.相比类别保留投影,能有效避免类别保留投影在基因表达数据降维上存在的矩阵奇异和过拟合问题.通过对真实基因表达数据进行数据可视化和分类识别,验证了方法的有效性.

关键词: 基因表达数据;高维小样本;类别保留投影;回归
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2016)04-0873-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.04.017

New Method of Dimensionality Reduction for Gene Expression Data Based on Sparse Class Preserving Projection

WANG Wen-jun
(*School of Computer Science and Engineering, Xidian University, Xi'an, Shaanxi* 710071, *China*)

Abstract: To solve the problem of the curse of dimensionality of gene expression data due to the characteristic of high dimension low sample size, a new method of dimensionality reduction for gene expression data, called sparse class preserving projection (SCPP) is proposed, by combining regression and class preserving projection(CPP). Compared to CPP, SCPP can avoid the problems of matrix singularity and over-fitting. Experiments are performed on gene expression data for visualization and sample classification, and the results confirm the effectiveness of the method.

Key words: gene expression data; high dimension and low sample size; class preserving projection; regression

1 引言

基于基因表达数据[1~3]的肿瘤分类[4~6]研究对癌症诊疗有着非常重要的意义.基因表达数据的高维小样本特性,成为传统模式分类研究的瓶颈.如何有效降低基因表达数据的维数,成为基因表达数据分类研究的关键问题之一.

为克服这一问题,已有学者尝试开发基因表达数据降维方法.这些方法主要包括基于基因选择的方法、基于非监督特征提取的方法和基于监督特征提取的方法:

(1)基于基因选择[7~14]的方法,这是目前基因表达数据降维的最主要的方法.基因选择通过选取差异显著基因可能会达到很高的分类正确率,但并没有考虑基因之间的关系.很多疾病并不单纯是由差异显著基因的改变造成的,而是由复杂调控机制的改变引起的,所以很多疾病易感基因在不同类别样本间的表达并没有显著差异,但基因选择很可能会丢失这些疾病易感基因.面对不同的肿瘤分类任务,各种基因选择算法并没有统一的标准,如果基因选择算法设计的不好,就可能丢失对分类有用的信息基因,从而影响分类性能.

(2)基于非监督特征提取的方法,包括主分量分析(PCA)[15,16]、独立分量分析(ICA)[17]、非负矩阵分解法(NMF)[18]和保局投影(LPP)[19]等.这些特征提取方法都是没有考虑分类信息的降维方法,降维后往往还需借助一些鉴别特征提取方法来提取有效的分类特征,或采用支持向量机(SVM)[20]等比较复杂的分类器来提高分类性能,从而增加了分类识别的复杂性.

(3)基于监督特征提取的方法,经典监督特征提取方法是线性鉴别分析(LDA)[21].相比基因选择和非监督特征提取方法,监督特征提取方法能避免基因选择带来的信息丢失问题,同时减轻分类器设计的负担.在基因表达数据的应用上,LDA主要是用于数据降维后

收稿日期:2014-07-23;修回日期:2014-12-22;责任编辑:孙瑶
基金项目:中央高校基本科研业务费专项资金(No. JB140310)

的鉴别特征提取,而没有直接用来实现高维基因表达数据的降维.这主要是由于 LDA 面对基因表达数据的高维小样本特性,存在计算复杂度高、矩阵奇异、过拟合和最优子空间维数受样本类别数限制等问题,使 LDA 作为基因表达数据的降维手段遇到了瓶颈.类别保留投影(Class Preserving Projection,CPP)[22]是 2012 年提出的一种新的监督特征提取方法,CPP 能有效解决最优子空间维数受样本类别数限制的问题,同时基于样本空间的鉴别特征提取[23]能大大降低特征提取的计算复杂度.但面对基因表达数据的高维小样本特性,CPP 依然存在矩阵奇异、过拟合等问题.

为克服类别保留投影方法的不足,本文提出一种基于稀疏类别保留投影的基因表达数据降维方法,将 CPP 方法和线性回归相结合,避免类别保留投影在基因表达数据降维上存在的矩阵奇异和过拟合问题,提高肿瘤基因表达数据分类的准确性和可靠性.

2 方法

给定 m 个训练样本的基因表达数据矩阵 $\boldsymbol{X}_{n\times m}$和样本类别属性集合 $\boldsymbol{C}=[c_1,c_2,\cdots,c_m]$.矩阵 $\boldsymbol{X}$ 的行代表基因,列代表组织样本(简称"样本"),其元素 $\boldsymbol{x}_{ij}$是基因 i 在样本 j 上的表达水平.每个样本对应一个 n 维的表达向量,即 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\in\mathbf{R}^n$,样本 $\boldsymbol{x}_i$的类别记为 c_i.找一个变换矩阵 $\boldsymbol{A}$,使这 m 个样本映射到 d 维空间中的 m 个点:$\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_m\in\mathbf{R}^d$,使得 $\boldsymbol{y}_i$代表 $\boldsymbol{x}_i$,这里 $\boldsymbol{y}_i=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i$.

2.1 类别保留投影 CPP

CPP 是 2012 年提出的一种鉴别特征提取方法,从两两样本的类别关系出发,样本的类别关系作为权重系数,构造目标函数,使同类的任意两样本的距离尽可能地小,而异类的任意两样本之间的距离尽可能地大.相比经典的线性鉴别分析方法(LDA),CPP 具有最优子空间维数不受样本类别数限制、计算复杂度低的优点.

设 $\boldsymbol{a}$ 是一个变换向量,样本 $\boldsymbol{x}_i$ 在 $\boldsymbol{a}$ 上的投影记为 $\boldsymbol{y}_i$,即 $\boldsymbol{y}_i=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i,i=1,\cdots,m$,CPP 的目标函数为:

$$\min\frac{\sum\limits_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2\boldsymbol{W}_{ij}^1}{\sum\limits_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2\boldsymbol{W}_{ij}^2}\tag{1}$$

其中,$\boldsymbol{W}_{ij}^1=\begin{cases}1, & \text{if } c_i=c_j\\ 0, & \text{else}\end{cases}$,$\boldsymbol{W}_{ij}^2=\begin{cases}1, & \text{if } c_i\neq c_j\\ 0, & \text{else}\end{cases}$.

把 $\boldsymbol{y}_i=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i,i=1,\cdots,m$ 代入式(1),通过简单的代数变换,目标函数可化简为

$$\frac{\frac{1}{2}\sum\limits_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2\boldsymbol{W}_{ij}^1}{\frac{1}{2}\sum\limits_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2\boldsymbol{W}_{ij}^2}=\frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^1\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}}\tag{2}$$

其中,$\boldsymbol{L}^k=\boldsymbol{D}^k-\boldsymbol{W}^k,k=1,2,\boldsymbol{D}_{ii}^k=\sum_j\boldsymbol{W}_{ij}^k$.

使目标函数(2)最小的变换向量 $\boldsymbol{a}$ 可通过求解以下的广义特征方程来获得:

$$\boldsymbol{X}\boldsymbol{L}^1\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}=\lambda\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}\tag{3}$$

广义特征方程(3)的最小特征值对应的特征向量就是最优变换向量 $\boldsymbol{a}$.广义特征方程的前 d 个最小特征值对应的特征向量 $\boldsymbol{a}_i(i=1,2,\cdots,l)$就构成了 CPP 的最优变换矩阵 $\boldsymbol{A}=(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_d)$.

对于基因表达数据而言,由于其高维小样本特性,CPP 容易出现数据堆积(data piling)而出现过学习,从而降低方法的推广能力.同时,由于 $n\gg m$,所以由 m 个样本计算的类间散布矩阵 $\boldsymbol{X}\boldsymbol{L}^2\boldsymbol{X}^{\mathrm{T}}$ 一定是严重奇异的,这些问题在 LDA 特征提取方法中同样存在.为解决这些问题,我们结合回归方法,提出稀疏类别保留投影 SCPP.

2.2 稀疏类别保留投影 SCPP

SCPP 将回归和 CPP 相结合,把广义特征值问题转化到回归框架,特征向量转化为回归系数,用 elastic net[24]获得回归系数的稀疏解,提高特征的可解释性.

步骤 1 通过求解以下的广义特征方程来获得训练样本的鉴别特征映射

$$\boldsymbol{L}^1\boldsymbol{y}=\lambda\boldsymbol{L}^2\boldsymbol{y}\tag{4}$$

广义特征方程的前 d 个最小特征值对应的特征向量就是训练样本的鉴别特征映射 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_d\}$.$\boldsymbol{Y}$ 为 $m\times d$ 维的矩阵,d 为特征维数.

步骤 2 获得稀疏变换矩阵

通过求解以下的回归优化问题获得稀疏的列向量 $\boldsymbol{a}_i$:

$$\arg\min_{\boldsymbol{a}_i}\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}_i-\boldsymbol{y}_i\|+\alpha\|\boldsymbol{a}_i\|_2+\beta\|\boldsymbol{a}_i\|_1\tag{5}$$

获得的 d 个稀疏向量 $\boldsymbol{a}_i$ 组成最终的稀疏变换矩阵 $\boldsymbol{A}=\{\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_d\}$.

步骤 3 通过稀疏变换矩阵实现数据降维

对于任一样本 $\boldsymbol{x}_{\mathrm{new}}$,求得其在 $\boldsymbol{A}$ 上的投影值为

$$\boldsymbol{y}_{\mathrm{new}}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{new}}\tag{6}$$

$\boldsymbol{x}_{\mathrm{new}}\in\mathbf{R}^n,\boldsymbol{y}_{\mathrm{new}}\in\mathbf{R}^d,d\ll n$,实现数据降维.

由于变换矩阵 $\boldsymbol{A}$ 是由稀疏列向量 $\boldsymbol{a}_i$ 组成的,所以样本 $\boldsymbol{x}_{\mathrm{new}}$的特征向量 $\boldsymbol{y}_{\mathrm{new}}$对应的每个特征都是少数基因线性组合的结果,从而使特征更具解释性.

3 实验

对 NCI 和 GCM 这两组真实基因表达数据进行实验研究,采用 SCPP 进行降维,并与 CPP 进行比较,从可视化效果和分类识别准确率以及特征基因选择方面验证 SCPP 的有效性.

NCI 数据[25] 该数据由美国癌症研究院(NCI)提供的来自 neuroblastoma 神经细胞和非霍吉金氏淋巴细胞肿瘤这两类样本的源基因表达数据,这是在 4 种病类

88 个人的人群中所采集的这些人的基因表达数据,即其基因空间的维数为2308,样本数为88,病类数为4,其中64 个样本的所属类别已知,各病类中的样本数分别为:23、8、12、21.

GCM 数据[26]　190 例不同癌症类型的组织样本的16063 个基因片段的表达情况. 该数据包含 14 种癌症类型.

两组实验数据的详细信息如表 1.

表 1　实验数据

数据集名称	基因数	样本数	样本类别数
NCI 数据	2308	64	4
GCM 数据	16063	190	14

3.1　可视化效果和分类准确率

分别采用 SCPP 方法和 CPP 方法实现数据降维后,实现数据在前三个特征(主分量)上的可视化,并在鉴别特征空间采用最近邻分类器进行样本分类识别,计算样本分类的正确率,并通过 5 重交叉验证分析方法的推广能力. 数据的可视化结果见图 1 和图 2,分类正确率曲线见图 3 和图 4. 图 3、图 4 中,横坐标表示降维后的特征维数,纵坐标表示采用5 重交叉验证和最近邻分类器获得的分类正确率. NCI 数据降维后的特征维数最高选到 49,GCM 数据降维后的特征维数最高选到 150.

从图 1、图 2 可以看出,SCPP 在前三个特征上的数据散布不会出现数据堆积现象,CPP 的数据堆积严重. 从图 3、图 4 可以看出,SCPP 的分类正确识别率要明显好于 CPP,而且 SCPP 的正确识别率比较平稳,随着特征维数的增加呈单调不减趋势,而 CPP 达到最高识别率后,随着特征维数的增加,识别率下降明显. 对于 NCI 数据,在特征维数为 3 时,SCPP 的正确识别率达到了最大值(达到了 98.44%),而 CPP 的最高正确识别率是 96.88%. 对于 GCM 数据,CPP 的最高正确识别率是 62.63%(此时的特征维数是 16),而 SCPP 在特征维数为 13 时,正确识别率已经达到了 65.79%,最高正确识别率可达到 66.32%. 这说明 SCPP 能有效避免 CPP 的过拟合问题,从而提高特征的推广能力.

3.2　特征相关基因选择

在特征变换向量上系数不为零的基因称为特征相关基因. 对 NCI 数据采用 SCPP 方法获得的特征中,只有 4 个特征有相关基因,见表 2.

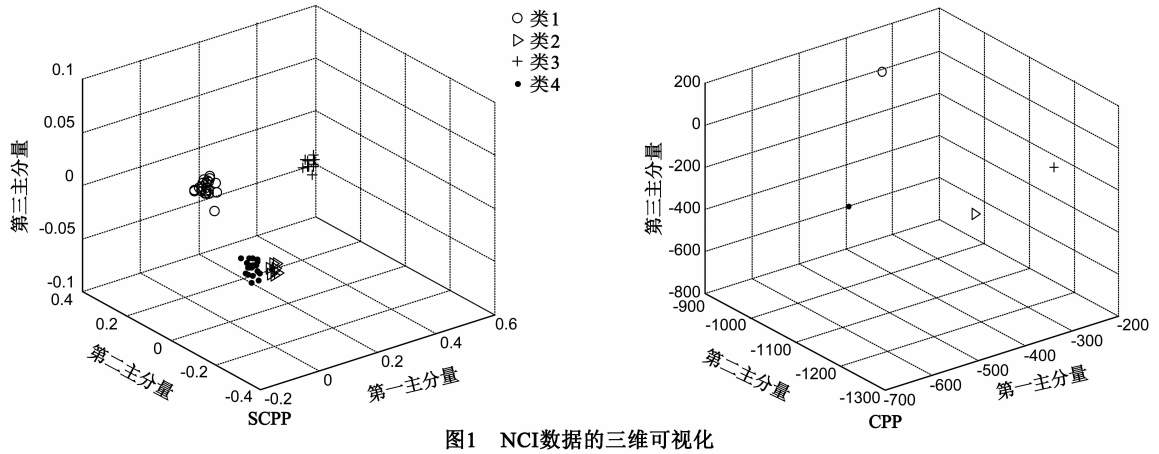

图1　NCI数据的三维可视化

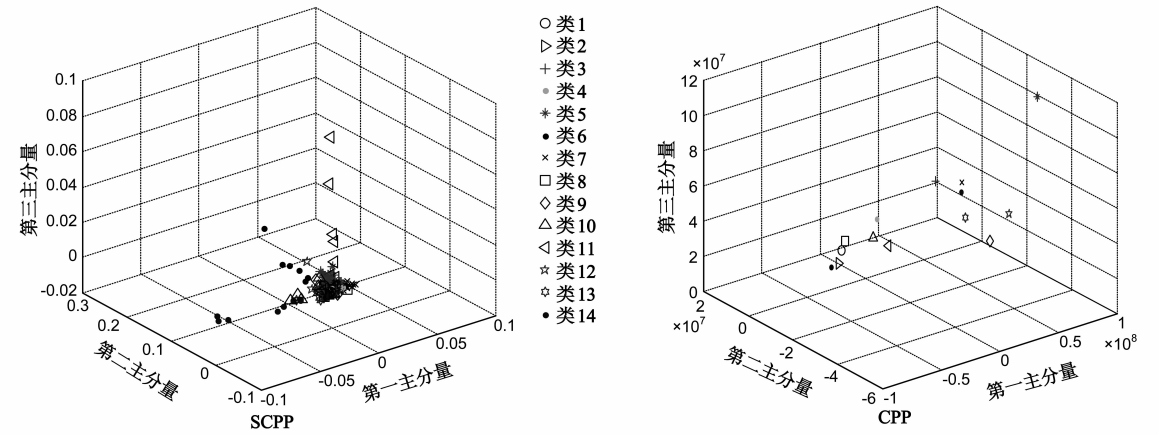

图2　GCM数据的三维可视化

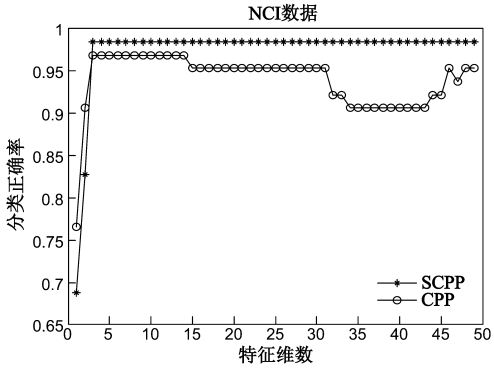

图3 NCI数据的分类正确率曲线

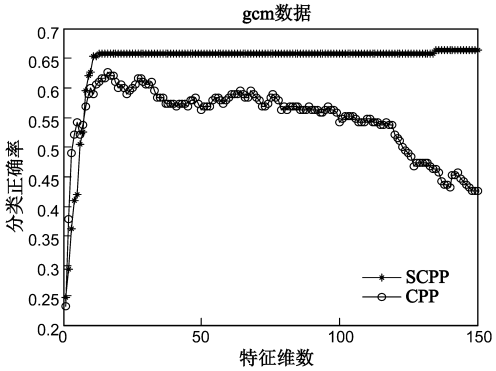

图4 GCM数据的分类正确率曲线

表 2 NCI 数据的特征相关基因

第 1 个特征	22 个相关基因： 153 **255** 469 761 849 857 867 1093 1207 1221 1283 **1389** 1536 **1601** 1662 1764 1867 1888 1908 2050 2144 2275
第 2 个特征	45 个相关基因： 94 107 123 246 **255** 365 384 477 521 532 533 545 589 607 667 719 731 758 846 865 881 937 998 1002 1036 1099 1142 1324 **1389** 1443 1453 1479 1522 1565 1606 1657 1778 1884 1932 1949 1956 1974 2127 2159 2162
第 3 个特征	33 个相关基因： 67 **174** **255** 426 483 509 603 624 655 742 828 879 910 951 1003 1030 1158 1316 1434 1576 **1601** 1626 1723 1738 1804 1924 1955 1980 2000 2083 2153 2157 2203
第 36 个特征	9 个相关基因： **174** 212 361 574 656 689 754 1291 1716

从表 2 可以看出，SCPP 从 2308 个基因中选出了 104 个不重复的特征相关基因. 这四个不同特征的特征相关基因及其个数几乎都不相同，只有少数基因在不同特征中重复出现. 如编号为 255 的基因跟前三个特征都相关，而基因 174、1389、1601 只与其中 2 个特征相关. 可见，SCPP 并不是象基因选择那样，只选出少数特征基因作为分类基因，而是不同的特征包含了许多不同的特征相关基因，故能更多地保留信息基因.

限于论文篇幅，GCM 数据的特征相关基因没有列出.

4 结论

本文将类别保留投影问题转化到回归框架，实现稀疏鉴别特征提取，克服了类别保留投影在基因表达数据降维上存在矩阵奇异、过拟合的问题. 采用稀疏类别保留投影实现基因表达数据降维，避免基因选择所带来的信息基因丢失，减轻分类器设计的负担，提高肿瘤基因表达数据分类的准确性和可靠性.

参考文献

[1] Rung J, Brazma A. Reuse of public genome-wide gene expression data[J]. Nature Reviews Genetics, 2013, 14(2): 89 – 99.

[2] 于攀，叶俊勇. 基于谱回归和核空间最近邻的基因表达数据分类[J]. 电子学报，2011，39(8)：1955 – 1960.
YU Pan, YE Jun-yong. Spectral regression and kernel space K-nearest neighbor for classification of gene expression data[J]. Acta Electronica Sinica, 2011, 39(8): 1955 – 1960. (in Chinese)

[3] Pham TD, Wells C, Crane DI. Analysis of microarray gene expression data[J]. Current Bioinformatics, 2006, 1(1): 37 – 53.

[4] Wang Z, Palade V. Fuzzy Models for High Dimensional Cancer Gene Expression Data Classification[D]. University of Oxford, 2013.

[5] Zhang YJ, Xuan JH, Clarke R, Ressom HW. Module-based breast cancer classification[J]. International Journal of Data Mining and Bioinformatics, 2013, 7(3): 284 – 302.

[6] 王年，庄振华，范益政，李学俊，王继. 癌症基因分类的 Laplace 谱方法[J]. 电子学报，2011，20(7)：1594 – 1597.
WANG Nian, ZHUANG Zhen-hua, FAN Yi-zheng, LI Xue-jun, WANG Ji. Classification of tumor gene expression data based on Laplacian spectra of graphs[J]. Acta Electronica Sinica, 2011, 20(7): 1594 – 1597. (in Chinese)

[7] Mao Z, Cai W, Shao X. Selecting significant genes by randomization test for cancer classification using gene expression data[J]. Journal of Biomedical Informatics, 2013, 46(4): 594 – 601.

[8] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines[J]. Mach Learn, 2002, 46(1 – 3): 389 – 422.

[9] Chen KH, Wang KJ, Tsai ML, et al. Gene selection for cancer identification: a decision tree model empowered by

particle swarm optimization algorithm[J]. BMC Bioinformatics,2014,15(1):Article ID 49.

[10] Cui Y,Zheng CH,Yang J,Sha W. Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data[J]. Computers in Biology and Medicine,2013,43(7):933 - 941.

[11] Ghosh S,Mitra S,Dattagupta R. Fuzzy clustering with biological knowledge for gene selection[J]. Applied Soft Computing,2014,16(1):102 - 111.

[12] Gusnanto A,Ploner A,Shuweihdi F,Pawitan Y. Partial least squares and logistic regression random-effects estimates for gene selection in supervised classification of gene expression data[J]. Journal of Biomedical Informatics,2013,46(4):697 - 709.

[13] Mohamad MS,Omatu S,Deris S,et al. An enhancement of binary particle swarm optimization for gene selection in classifying cancer classes[J]. Algorithms for Molecular Biology,2013,8(1):Article ID 15.

[14] Zhang HY,Wang HY,Dai ZJ,et al. Improving accuracy for cancer classification with a new algorithm for genes selection[J]. BMC Bioinformatics,2012,13(1):Article ID 298.

[15] Lee D,Lee W,Lee Y,Pawitan Y. Super-sparse principal component analyses for high-throughput genomic data[J]. BMC Bioinformatics,2010,11(1):Article ID 296.

[16] Liu JX,Wang YT,Zheng CH,et al. Robust PCA based method for discovering differentially expressed genes[J]. BMC Bioinformatics,2013,14(S8):Article ID S3.

[17] Huang DS,Zheng CH. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data[J]. Bioinformatics,2006,22(15):1855 - 1862.

[18] Zheng CH,Ng TY,Zhang L,et al. Tumor classification based on non-negative matrix factorization using gene expression data[J]. IEEE Transactions on Nanobioscience,2011,10(2):86 - 93.

[19] He XF,Niyogi P. Locality preserving projections[A]. Advances in Neural Information Processing Systems[C]. USA:MIT Press,2004,Vol 16. 153 - 160.

[20] Liu J,Li SC,Luo X. Iterative reweighted noninteger norm regularizing SVM for gene expression data classification[A]. Computational and Mathematical Methods in Medicine[C]. USA:Hindawi Publishing Corporation,2013. Article ID 768404.

[21] Paliwal KK,Sharma A. Improved direct LDA and its application to DNA microarray gene expression data[J]. Pattern Recognition Letters,2010,31(16):2489 - 2492.

[22] 王文俊. 基于类别保留投影的基因表达数据特征提取新方法[J]. 电子学报,2012,40(2):358 - 364.
WANG Wen-jun. New method of feature extraction for gene expression data based on class preserving projection[J]. Acta Electronica Sinica,2012,40(2):358 - 364. (in Chinese)

[23] Wang WJ. Sample-space-based feature extraction and class preserving projection for gene expression data[J]. International Journal of Data Mining and Bioinformatics,2013,8(2):224 - 246.

[24] Zou H,Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society Series B-Statistical Methodology,2005,67(2):301 - 320.

[25] Khan J,Wei JS,Ringner M,et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks[J]. Nature Medicine,2001,7(6):673 - 679.

[26] Ramaswamy S,Tamayo P,Rifkin R,et al. Multiclass cancer diagnosis using tumor gene expression signatures[J]. Proceedings of the National Academy of Sciences of the United States of America,2001,98(26):15149 - 15154.

作者简介

王文俊　女,1980 年 8 月出生,山西平遥人,副教授、硕士生导师. 2003 年、2006 年和 2011 年在西安电子科技大学分别获得工学学士、工学硕士和工学博士学位. 2006 年至今在西安电子科技大学计算机学院从事教学科研工作,主要研究方向为模式识别、生物信息处理等. E-mail:xidianwwj219@163.com