基于偏最小二乘与极大间距准则的微阵列分类*

胡 煜[1]，阳文辉[2]
（1. 广东工贸职业技术学院，广东 广州 510510；
2. 中山大学数学与计算科学学院，广东 广州 510275）

摘 要：针对基因表达数据中的高维小样本问题，提出了一种两阶段的识别框架："偏最小二乘法（PLS）+极大间距准则（MMC）"。该方法首先使用PLS算法提取出带有分类信息的特征，然后使用MMC准则对样本进行分类。在六个公共的基因数据库上与一些常见的基因分类方法相比较，结果显示了该方法对基于基因表达数据的肿瘤分类有效且稳定。

关键词：微阵列数据；基因选择；特征提取；偏最小二乘；极大间距准则
中图分类号：TP391.41 **文献标识码**：A **文章编号**：0529-6579（2008）04-0005-06

在现代肿瘤学中，病理学家的角色首先必须识别肿瘤的类型或肿瘤的起源细胞，这一步在肿瘤的治疗与预防中有非常重要的作用。当前肿瘤的主要分类方法是利用肿瘤或周边组织的形态学进行模式识别，然而这种方法在肿瘤早期很难正确地分类。癌症本质上是由于基因的异常表达而引起恶性细胞的增长，所以最直接的方法是对样本的基因表达信息进行分析。20世纪90年代发展的微阵列技术能同时测定上万个基因的表达水平，该项技术期待能实现肿瘤早期的精确诊断，从而为肿瘤的治疗做出重要贡献。有关微阵列的详情可参见文［1-2］。

利用微阵列技术可以检测出不同样本的基因表达水平。在有监督环境中，部分样本的肿瘤类型是已知的，我们的目标是利用新样本的表达水平来预测其肿瘤类型。近年来，基于样本的基因表达数据，提出了很多分类方法。主要方法有线性判别方法（LDA）[1,3]，提升法（Boosting）[4]，logistic回归判别法[5-6]，贝叶斯（Bayesian）方法[7]，支持向量机（SVM）[3]等。极大间距准则（MMC）[8]是一个效率高、鲁棒性好的分类方法，在人脸识别和基因分类中有广泛的推广和应用[9-10]。与传统的线性判别方法相比，MMC可以避免小样本问题。在本文中，将引入MMC来实现基因的分类。

微阵列数据的特点是样本数目较少（一般是几十或几百个），而基因数目非常多（几千甚至上万个），这种数据通常称为高维小样本数据。这种数据为随后的统计分析工作带来了极大的挑战。目前，最主要的方法是在分类之前对数据进行预处理。处理的方法有基因选择和特征提取。基因选择在数据预处理中是一个非常重要的步骤，处理的好坏直接影响到后面的分类结果。在两类问题中，常用的基因选择方法有 t-统计量[5]，Wilcoxon统计量[4]等。在多类问题中，常用的基因选择方法有成对 t-统计量[6]，BSS/WSS准则[1]，软阈值处理等。文［3］对各种基因选择方法的结果进行了比较。

另外，对高维数据的特征提取也非常重要[11]。尽管基因的数目非常多，但只有少量的特征对分类是有效的。特征提取可以把高维空间的样本投影到低维空间中。主成分分析（PCA）和偏最小二乘法（PLS）是两个主要的特征提取方法。PCA的目标是找一个线性投影，在投影方向上样本的方差最大。PCA是一种无监督投影方法，没有考虑类的标记，从而不可避免的会对后面的分类方法有一些损失。与PCA方法不同，PLS考虑类的标记，是一种有效的有监督投影方法。有关PLS的理论可参考文［12-13］。

在本文中，基于样本的基因表达数据，我们提出了一个两阶段的肿瘤分类方法："PLS+MMC"。该方法首先使用PLS算法提取出带有分类信息的特征，然后使用MMC准则对样本进行分类。最后，在6个公共的基因数据库上与一些常见的基因

* **收稿日期**：2007-12-01
基金项目：国家自然科学基金资助项目（60575004，10771220）；广东省自然科学基金资助项目（0510181）；教育部"新世纪优秀人才支持计划"资助项目（NCET-04-0791）
作者简介：胡煜（1964年生），男，讲师；E-mail：cgzhy163@163.com

分类方法相比较，检验该方法的性能。

1 算法的提出

当预测变量个数超过样本数目时，很多传统的统计分类方法都遇到了障碍。在基因表达数据中，样本的数目远远小于基因个数。因此，找一种能够处理高维小样本数据的分类方法是至关重要的。在这一节中，将介绍一种 PLS 特征提取与 MMC 分类器相结合的算法。

1.1 基因选择

肿瘤患者仅仅在少量基因中出现异常表达，大量基因在微阵列数据中的表达水平是不显著的，因此在微阵列分析之前必须进行基因选择。文［3］中对 Wilcoxon 统计量，BSS/WSS 准则，软阈值处理等基因选择方法的结果进行了比较，BSS/WSS 准则有更好的表现。因此，在本文中引入 BSS/WSS 准则来选择基因，这个准则在文［1］中有详细描述。对第 j 个基因的 BSS/WSS 得分为

$$F_j = \frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k \mathrm{I}(\mathrm{y_i} = \mathrm{k})(\bar{\mathrm{g}}_{\mathrm{j}}^{(\mathrm{k})} - \bar{\mathrm{g}}_{\mathrm{j}})^2}{\sum_i \sum_k \mathrm{I}(\mathrm{y_i} = \mathrm{k})(\mathrm{x_{ji}} - \bar{\mathrm{g}}_{\mathrm{j}}^{(\mathrm{k})})^2}$$

其中 $\bar{g}_j$ 表示第 j 个基因在所有样本中的平均表达水平，$\bar{g}_j^{(k)}$ 表示第 j 个基因在第 k 类样本中的平均表达水平，x_{ji} 是第 j 个基因在第 i 个样本中的表达水平，y_i 是样本的类标记。按照每个基因的得分，从大到小排列，得分高的基因被选择，得分低的基因认为是不显著的，因而被淘汰。

1.2 PLS 特征提取

通过 BSS/WSS 得分，假设得分最高的前 p 个基因被选择。令 $X = [\mathrm{x}_1, \cdots, \mathrm{x}_i, \cdots, \mathrm{x}_N]$ 为基因选择后的数据矩阵，大小为 $p \times N$，其中 N 是样本数目，x_i 是第 i 个样本在 p 个基因中的表达水平。

在 PCA 里，正交的特征投影集是通过依次最大化预测变量的线性组合得到的，如求第 k 个特征投影：

$$v_k = \underset{v^{\mathrm{T}}v=1}{\operatorname{argmaxvar}}(X^{\mathrm{T}}v)$$

$$\text{受限于} \quad v^{\mathrm{T}}S_t v_j = 0, \quad 1 \leqslant j < k,$$

其中 $S_t = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)(x_i - m)^{\mathrm{T}}$，$m$ 是所有样本的平均。主成分分析的主要目的是要提取隐藏在矩阵 X 中的主要信息，降低数据噪声，从而改善预测模型的质量。但是，主成分分析仍然有一定的缺陷，当一些有用变量的相关性很小时，在选取主成分时就很容易把它们漏掉，使得最终的预测模型可靠性下降。

PLS 是一种新型的多元统计数据分析方法，它于 1983 年由伍德和阿巴诺等首次提出。近几十年来，它在理论、方法和应用方面都得到了迅速的发展。PLS 能同时实现回归建模（多元线性回归）、数据结构简化（主成分分析）以及两组变量之间的相关性分析。PLS 回归能有效地解决 PCA 出现的问题，利用大量的连续预测变量来恰当地预测一个或多个响应。它采用对预测变量 X 和响应变量 Y 都进行分解的方法，从变量 X 和 Y 中同时提取成分（通常称为因子），再将因子按照它们之间的相关性从大到小排列。假设预测变量 X 和响应变量 Y 都已经中心化，从变量 X 中提取的第 k 个投影为：

$$v_k = \underset{v^{\mathrm{T}}v=1, c^{\mathrm{T}}c=1}{\operatorname{argmax}} \operatorname{cov}^2(X^{\mathrm{T}}v, Y^{\mathrm{T}}c)$$

$$\text{受限于} \quad v^{\mathrm{T}}S_t v_j = 0, \quad c^{\mathrm{T}}YY^{\mathrm{T}}c_j = 0, 1 \leqslant j < k$$

如果基因分类是一个两类问题，则响应变量 Y 是一个二值响应（0 或 1）。因为 PLS 回归不要求任何分布的假定，所以，此时 Y 可以看成一个连续响应变量。然而当基因分类是一个多类问题时，Y 不能看成一个连续响应变量。这种情况下，可以用哑元编码的方式来实现。假如基因分类中包含 K 种类型，则多类随机变量 Y 转换成一个 K －维的随机变量：

$$y \in \{0,1\}^K$$

$$\begin{cases} y_{ki} = 1 & \text{如果 } Y_i = k, \\ y_{ki} = 0 & \text{否则}, \end{cases}$$

其中 $y_i = (y_{1i}, \ldots, y_{Ki})^{\mathrm{T}}$ 表示第 i 个样本响应的实现。转换后的响应矩阵 Y 的大小为 $K \times N$。

作为一个多元线性回归方法，PLS 回归的主要目的是要建立一个线性模型：

$$Y = BX + E$$

其中 $Y \in \{0,1\}^{K \times N}$ 是响应矩阵，$X \in \mathbb{R}^{p \times N}$ 是预测矩阵，$B \in \mathbb{R}^{K \times p}$ 是回归系数矩阵，E 为噪音校正模型，与 Y 具有相同的维数。在通常情况下，变量 X 和 Y 被标准化后再用于计算，即中心化后再除以标准偏差。但在我们的应用中，并不需要求出回归系数矩阵 B，只需要求出预测矩阵 X 的前 d 个加权系数向量 $v_i (i = 1, \cdots, d)$ 组成的加权矩阵 V。算法 1 描述了在我们应用中的迭代 PLS 算法。

利用加权矩阵 V，可以建立一个映射：

$$V: x \in \mathbb{R}^p \rightarrow z = V^{\mathrm{T}}x \in \mathbb{R}^d,$$

实现基因数据的特征提取。提取后基因数据矩阵变为。

$$Z = V^{\mathrm{T}}X \in \mathbb{R}^{d \times N}$$

算法 1. PLS 迭代算法

输入：预测矩阵 $X \in \mathbb{R}^{p \times N}$，响应矩阵 $Y \in \{0,1\}^{K \times N}$，PLS 成分个数 d.

输出：加权系数矩阵 $V \in \mathbb{R}^{p \times d}$

初始化：$X_{(1)} = X, Y_{(1)} = Y$

FOR $k = 1$ to d
　u^{T} 为 $Y_{(k)}$ 的第一行
　DO
　　$v_k = Xu/(u^{\mathrm{T}}u)$ 并把 v_k 单位化
　　$t = X^{\mathrm{T}}v_k, c = Yt/(t^{\mathrm{T}}t)$ 并把 c 单位化
　　$u = Y^{\mathrm{T}}c$
　Until u 收敛
　$s = Xt/(t^{\mathrm{T}}t), b = u^{\mathrm{T}}t/(t^{\mathrm{T}}t)$
　计算残差矩阵 $X_{k+1} = X_k - st^{\mathrm{T}}, Y_{k+1} = Y_k - bct^{\mathrm{T}}$
END FOR

1.3　极大间距准则

经过 PLS 特征提取后的数据矩阵 $Z \in \mathbb{R}^{d\times N}$，一般地，$d$ 远远小于 p。在较低维空间上，很多统计方法都可以实现分类。线性判别分析（LDA）是一种广泛应用的分类方法，包括微阵列数据分类、人脸识别、文本分类等。它的目标是找一个投影矩阵 $W \in \mathbb{R}^{q\times d}$，使目标矩阵

$$J(W) = \frac{|W\tilde{S}_b W^{\mathrm{T}}|}{|W\tilde{S}_w W^{\mathrm{T}}|}$$

最大化，这个准则又称为 Fisher 商，这里 $\tilde{S}_b$ 为定义在数据矩阵 Z 上的类间散度矩阵，$\tilde{S}_w$ 为类内散度矩阵。但是，当样本数较少，或者 PLS 降维不是足够低的时候，LDA 仍然会有小样本问题。因此，找一个可以避免小样本问题的分类方法就变得尤为重要。

极大间距准则（MMC）[8]是一个效率高、鲁棒性好的分类方法。跟 LDA 类似，MMC 也是找一个投影矩阵 $W \in \mathbb{R}^{q\times d}$，使 $|W\tilde{S}_b W^{\mathrm{T}}|$ 最大，且 $|W\tilde{S}_w W^{\mathrm{T}}|$ 最小。MMC 准则是以差的形式代替 LDA 里商的形式，其准则为：

$$W_{\mathrm{MMC}} = \underset{W\in\mathbb{R}^{q\times d}, WW^{\mathrm{T}}=I_q}{\operatorname{argmax}} \operatorname{Tr}(W(\tilde{S}_b - \tilde{S}_w)W^{\mathrm{T}})$$

其中 $\operatorname{Tr}(A)$ 代表矩阵 A 的迹，I_q 为 $q \times q$ 的单位矩阵。因为 MMC 准则不需要考虑类内散度矩阵 $\tilde{S}_w$ 的奇异性，因此，MMC 准则可以避免小样本问题。

在 MMC 准则里，投影矩阵 W 的计算非常容易，只需对 $\tilde{S}_b - \tilde{S}_w$ 进行特征分解即可。在我们的算法中，投影个数 q 为矩阵 $(\tilde{S}_b - \tilde{S}_w)$ 的非零特征值的个数，详细算法见（算法 2）。

算法 2.　极大间距准则算法

输入：特征提取后的数据矩阵 $Z \in \mathbb{R}^{d\times N}$，响应矩阵 Y

输出：投影矩阵 $W \in \mathbb{R}^{q\times d}$

1. 计算散度矩阵 $\tilde{S}_b$ 与 $\tilde{S}_w$
2. 对矩阵 $\tilde{S}_b - \tilde{S}_w$ 进行特征分解：$\tilde{S}_b - \tilde{S}_w = UAU^{\mathrm{T}}$
3. 从特征向量矩阵 U 中选取非 0 特征值所对应的特征向量，构造投影矩阵 W：

$$W = U(:, \operatorname{diag}(A) > 0)$$

1.4　算法流程

根据以上叙述，整个算法的流程如下：

第一步（预处理）　在不同实验条件下，基因表达的差异比较大，需要对原始基因数据进行归一化处理。为防止由个别基因主导，在本文中，先把实验所得到的数据进行以 10 为底的对数变换，然后进行标准化处理，处理后的样本平均值为零，标准差为 1。

第二步（基因选择）　标准化的基因数据中，对每一个基因求出 BSS/WSS 得分，按照从大到小排序，选出前 p 个基因。

第三步（特征提取）　基因选择后的数据为 $X \in \mathbb{R}^{p\times N}$，利用 PLS 迭代算法求出前 d 个加权系数向量组成加权矩阵 $V \in \mathbb{R}^{p\times d}$，然后建立映射 V：$\mathbb{R}^p \to \mathbb{R}^d$ 实现特征提取。

第四步（MMC 降维）　特征提取后的数据为 $Z \in \mathbb{R}^{d\times N}$，利用极大间距准则算法求出投影矩阵 $W \in \mathbb{R}^{q\times d}$，由投影矩阵建立映射 W:$\mathbb{R}^d \to \mathbb{R}^q$ 实现特征投影。完成投影后的 N 个训练样本已经降维到较低的 $\mathbb{R}^q$ 空间中。

第五步（预测）　给出一个新的样本，选择出指定的 p 个基因，然后通过加权矩阵 V 和投影矩阵 W 降维到 $\mathbb{R}^q$ 空间。在 q - 维空间中，新样本利用最近邻准则找到与之最近的训练样本，这个训练样本的类标就是新样本预测的类标。

2　实验结果与分析

2.1　数据集

在实验中共用到 6 个公共基因数据库：（1）白血病（Leukemia），共包含 72 例白血病人，其中有 47 例为急性淋巴白血病（ALL），25 例为急性骨髓白血病（AML）；（2）淋巴瘤（Lymphoma），共包含 62 例病人，其中 42 例为扩散的大的 B 型细胞淋巴瘤（DLBL），9 例为卵泡淋巴瘤（FL），11 例为慢性淋巴白血病（CLL）；（3）急性淋巴白血病（Acute Lymphoblastic Leukemia，ALL）；（4）小圆形蓝色细胞瘤（Small，Round Blue Cell Tumors，SRBCT）；（5）前列腺癌（Prostate），包含 52 例前列腺癌症病人样本和 50 例正常组织样本；（6）脑癌（Brain）。

这些数据集的汇总信息显示在表 1 中。

表1 实验中使用的公共基因数据集摘要

Tab. 1 Summary of the Data Sets Used in the Experiments

数据集	出版文献	样本个数	基因个数	类数	响应
Leukemia	Golub 等[1]	72	3571	2	白血病的子类型
Lymphoma	Alizadeh 等[2]	62	4026	3	子类型
ALL	Yeoh 等[14]	248	12625	6	子类型
SRBCT	Khan 等[15]	63	2308	4	子类型
Prostate	Singh 等[16]	102	6033	2	肿瘤/正常组织
Brain	Pomeroy 等[17]	42	5579	5	子类型

每个数据集的样本随机分成两部分，其中第一部分为2/3的训练样本，第二部分为1/3的检验样本。训练样本用于实现基因选择，特征提取以及分类器的拟合。用设计好的分类器来预测检验样本。识别率为预测准确的检验样本与整个检验样本的比例。为减少可变性，把数据集分割成训练集和检验集的过程重复200次，预测精度为这些识别率的平均。

2.2 实验结果

这一节包括2个实验，第一个实验考虑PLS + MMC算法中的参数调整对实验结果的影响，第二个实验是把PLS + MMC算法与一些经典方法相比较，检验该方法的性能。

实验1：参数选择。在PLS + MMC算法流程中，包含2个参数，第一个参数是基因个数的选择，另一个参数是PLS算法中成分个数的确定。在这个实验中，以SRBCT基因数据库为例来说明。在这个数据库中分别选择了50、100、200、300、500、1000个最显著基因。在不同基因个数条件下，分别选择30、50、70、100个PLS成分。实验结果显示在表2中。

表2 不同参数值条件下的实验结果

Tab. 2 The experimental results on different parametric values

/%

PLS特征提取数目 d	基因选择个数 p			
	30	50	70	100
50	92.32	92.32	92.08	91.79
100	96.90	96.90	96.67	96.61
200	98.93	98.99	98.99	98.81
300	98.87	99.17	99.17	99.11
500	98.81	98.87	98.81	98.81
1000	98.69	98.69	98.75	98.87

从表2中可以看出，当基因选择个数较少时，如选择50个基因，识别率不高，但随着基因个数的增加，识别率趋向于稳定。另外，也可以看出，我们的方法对PLS特征的数目不是特别敏感，只要选到一个合适的大小，识别率的差别不是很明显。从这个实验结果看出，本方法对参数不是特别敏感，基因个数为200～500，PLS特征个数为30～100时就可以取得较为满意的结果。

实验2：方法比较。在这个实验中通过与一些常见的分类算法相比较，检验PLS + MMC的性能，包括基于LDA的两个经典线性判别方法：PCA + LDA[18]与DLDA[19]，在文[1]中提到的两个性能较好的基因分类方法：k最近邻（k - NN），对角LDA（Diagonal LDA），以及与本文相关的分类方法MMC[8]相比较。

每个分类器都进行了基因选择，按BSS/WSS准则选择出200个基因。在文[1]中，Dudoit等认为从广泛的微阵列基因表达数据库来看，选择200个基因是一个合理的值。在PLS + MMC中，取PLS成分的个数为30来计算，得到的实验结果显示在表3中。

表3 6种分类器在6个数据集上的识别率比较

Tab. 3 Comparison of classification accuracy tate for six classifiers on six data sets

/%

数据集\方法	PLS + MMC	PCA + LDA	MMC	DLDA	k - NN	Diagonal LDA
Leukemia	95.73	61.98	96.17	77.29	96.25	85.16
Lymphoma	100.0	99.86	100.0	100.0	99.90	98.43
ALL	97.67	89.87	97.52	96.80	95.05	97.65
SRBCT	98.93	75.29	97.81	88.76	98.57	35.21
Prostate	87.65	90.51	86.44	77.91	85.21	87.24
Brain	93.84	93.57	93.57	92.64	92.71	93.50

从表3可以看出，PLS + MMC算法的性能最好，也最稳定。在Leukemia、Lymphoma、ALL和Brain四个数据库上，PLS + MMC与MMC无显著差异，但在SRBCT和Prostate上优势很明显。PLS把

样本从 200 维降到 30 维，从整体上来看，PLS 步骤改善了分类性能。PCA + LDA、DLDA 和 Diagonal LDA 三个分类器的性能不稳定，如 PCA + LDA 在 Leukemia 库上，DLDA 在 Leukemia 和 Prostate 库上，Diagonal LDA 在 SRBCT 库上，识别率非常低。综上所述，在基于基因表达数据的肿瘤分类中，PLS + MMC 算法是非常有效的。

3　结　语

对微阵列数据分类预测的目标是期望在恶性肿瘤早期阶段有一个精确的诊断，从而能为肿瘤的预防与治疗起到非常重要的作用。基因表达数据具有高维特征，设计的分类算法必须有处理这种数据的能力。在本文中，结合偏最小二乘与极大间距准则，提出了一种新的两阶段分类方法："PLS + MMC"。该方法至少有以下优势：（1）不受类的数目限制。可以处理两类问题，也可以处理多类问题；（2）不需要考虑数据的维数和样本数目的限制，克服了经典 LDA 的局限；（3）只需较少的特征（30 个左右）就能达到较高的识别精度。在六个公共基因数据上的实验结果显示了我们的算法有很好的判别能力和稳定性。

致谢：感谢戴道清教授对本文的有益指导和建议。

参考文献：

[1] DUDOIT S, FRIDLYAND J, SPEED T. Comparison of discrimination methods for the classification of tumors using gene expression data [J]. Journal of the American Statistical Association, 2002, 97: 77 - 87.

[2] ALIZADEH A A, EISEN M B, et al. Distinct types of diffuse large B-cell Lymphoma identified by gene expression profiling [J]. Nature, 2000, 40: 503 - 511.

[3] LEE J W, LEE J B, PARK M, SONG S H. An Extensive Comparison of Recent Classification Tools Applied to Microarray Data [J]. Computational Statistics & Data Analysis, 2005, 48: 869 - 885.

[4] DETTLING M. BagBoosting for tumor classification with gene expression data [J]. Bioinformatics, 2004, 20 (18): 3583 - 3593.

[5] NGUYEEN D V, ROCKE D M. Tumor classification by partial least squares using Microarray gene expression data [J]. Bioinformatics, 2002, 18: 39 - 50.

[6] NGUYEEN D V, ROCKE D M. Multi - class cancer classification via partial least squares with gene expression profiles [J]. Bioinformatics, 2002, 18: 1216 - 1226.

[7] ZHOU X B, WANG X D, DOUGHERTY E R. A Bayesian approach to nonlinear probit gene selection and classification [J]. Journal of the Franklin Institute, 2004, 341 (1 - 2): 137 - 156.

[8] LI H, JIANG T, ZHANG K. Efficient and robust feature extraction by maximum margin criterion [J]. IEEE Transactions on Neural Networks, 2006, 17(1): 157 - 165.

[9] YANG W H, DAI D Q, YAN H. Generalized discriminant analysis for tumor classification with gene expression data [C]. Proceedings of the Fifth International Conference on Machine Learning and Cybernetics (ICMLC' 06), 2006: 4322 - 4327.

[10] YANG W H, DAI D Q, YAN H. Feature extraction and uncorrelated discriminant analysis for high - dimensional data [J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(5): 601 - 614.

[11] DAI D Q, YAN H. Matrix decomposition for feature generation from high dimensional data, in "Pattern Recognition Theory and Application" [M]. Nova Science Publishers, 2007.

[12] HELLAND I S. On the structure of partial least squares [J]. Communications in Statistics-Simulation and Computation, 1988, 17: 581 - 607.

[13] HASTIE T, TIBSHIRANI R, FRIEDMAN J H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction [M]. Springer, 2001.

[14] YEOH E J, ROSS M E, SHURTLEFF S, et al. Classification, subtype discovery, and prediction of outcome in pediatric Lymphoblastic Leukemia by gene expression profiling [J]. Cancer Cell, 2002, 1: 133 - 143.

[15] KHAN J, WEI J, JUN S, et al. Classification and diagnostic prediction of cancers using expression profiling and artificial neural networks [J]. Nature Medicine, 2001, 7: 673 - 679.

[16] SINGH D, FEBBO P G, ROSS K, et al. Gene expression correlates of clinical prostate cancer behavior [J]. Cancer Cell, 2002, 1(2): 203 - 209.

[17] POMEROY S, TAMAYO P, GAASENBEEK M, et al. Prediction of central nervous system embryonal tumor outcome based on gene expression [J]. Nature, 2002, 415: 436 - 442.

[18] BELHUMEUR P N, HESPANHA J P, KRIGMAN D J. Eigenfaces vs Fisherface: recognition using class specific linear projection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19 (7): 711 - 720.

[19] YU H, YANG J. A direct LDA algorithm for high - dimensional data with application to face recognition [J]. Pattern Recognition, 2001, 34(10): 2067 - 2070.

Tumor Classification by Partial Least Squares and Maximal Margin Criterion Using Gene Expression Data

HU Yu[1], *YANG Wen-hui*[2]

(1. Guangdong Vocational College of Industry & Commerce, Guangzhou 510510, China;
2. Department of Mathematics, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: In order to deal with the high-dimensional and small sample size problem of gene expression data, a new two-phrase framework based on PLS and MMC is developed. this procedure involves feature extraction using partial least squares (PLS) and classification using maximal margin criterion (MMC). The proposed method is applied to six public microarray data sets. Experimental results demonstrate that this method is an effective and stable discrimination approach for tumor classification with gene expression data.

Key words: microarray data; gene selection; feature extraction; partial least squares; maximal margin criterion

(上接第4页)

参考文献:

[1] DAVID G, JOURNE J L. A boundedness criterion for generalized Calderon – Zygmund operators[J]. Ann of Math, 1984,120: 371 – 397.

[2] DENG D G, HAN Y S. *T*1 Theorem for Besov and Triebel – Lizorkin spaces[J]. Science in China Ser A Math,2004,34: 657 – 664.

[3] HAN Y S, LU S Z, YANG, D C. Inhomogeneous Besov and Triebel – Lizorkin space on spaces of homogeneous type[J]. Approx Theory Appl, 1999,15: 37 – 65.

[4] HAN Y S, SAWYER E T. Littlewood – Paley theory on spaces of homogeneous type and classical function spaces [J]. Mem Amer Math Soc, 1994,110: 1 – 126.

[5] YANG D C. *T*1 Theorem on Besov and Triebel – Lizorkin spaces on spaces of homogeneous type and their applications[J]. Zeitschriftfur Analysis ihre Anwendungen, 2003, 22: 53 – 72.

[6] DENG D G, HAN Y S, YANG D C. Inhomogeneous Plancherel – Polya type inequanlity on spaces of homogeneous type and their applications[J]. Communications in Contemporary Mathematics, 2004, 6: 221 – 243.

[7] HAN Y S. Plancherel – Polya type inequality on spaces of homogeneous type and its applications[J]. Proc Amer Math Soc, 1998,126: 3315 – 3327.

[8] CHRIST M. A $T(b)$ theorem with remarks on analytic capacity and the Cauchy integral[J]. Colloq Math, 1990,60/61: 601 – 628.

[9] HAN Y S, LU S Z, YANG D C. Inhomogeneous discrete Calderon reproducing formulas for spaces of homogeneous type[J]. J Fourier Anal Appl, 2001,7: 571 – 600.

[10] FEFERMAN C, STEIN E M. Some maximal inequalities [J]. Amer J Math, 1971,93: 107 – 115.

[11] FRAZIER M, JAWERTH B. A discrete transform and decompositions of distribution spaces[J]. J Funct Anal, 1990,93: 34 – 170.

*T*1 Theorem for Inhomogeneous Besov and Triebel – Lizorkin Spaces

HAN Yan-chang

(Department of Mathematics, South China Normal University, Guangzhou 510631, China)

Abstract: The *T*1 theorems for the inhomogeneous Besov and Triebel – Lizorkin spaces are established by the discrete Calderon type reproducing formula and the Plancherel – Polya characterization for the inhomogeneous Besov and Triebel – Lizorkin spaces on space of homogeneous type. These results are new even for $\mathbb{R}^d$.

Key words: *T*1 theorem; Besov and Triebel – Lizorkin spaces; Calderon reproducing formula; Plancherel – Polya inequalities