基于惩罚高斯混合模型的微阵列基因表达数据分析*

石 玉
（中山大学数学与计算科学学院，广东 广州 510275）

摘 要：随着现代生物技术的发展，基于基因表达数据的肿瘤分型诊断已成为 DNA 微阵列的重要应用领域。提出一种基于基因表达数据的肿瘤分型诊断新方法，并在理论上给出模型解释。该方法通过对高斯混合模型加上一个 L_1 惩罚实现了肿瘤分类和信息基因选择的有机结合，从而用较少的变量达到更高的识别率。实验结果显示，无论是在模拟数据中还是五个微阵列数据集中，提出的方法都是高效稳定的。

关键词：微阵列数据；肿瘤诊断；基因选择；混合高斯模型；L_1 惩罚
中图分类号：TP391 **文献标识码**：A **文章编号**：0529 - 6579（2009）03 - 0001 - 07

Penalized Gaussian Mixture Model for the Analysis of Microarray Gene Expression Data

SHI Yu
(Department of Mathematics, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: The diagnosis of cancer type based on gene expression profile is an important application field of DNA microarrays. In this paper, A penalized Gaussian mixture model is proposed by an L_1 penalty, which makes a good conjunction with cancer diagnosis and gene selection, achieve better accuracy by less variable deal. Some theoretic analysis is given for a simple but clear explanation about the model in detail. Simulate and real data examples show that the proposed method not only have high classification accuracy but also enjoy stable performance in various cases.

Key words: Microarray; Cancer diagnosis; Gene selection; Gaussian mixture model; L_1 penalty

随着人类基因组草图绘就的完成，人类基因组研究计划（Human Genome Project，HGP）进入到了后基因组时代。后基因组时代研究的重点由基因序列研究上升为基因功能研究。20 世纪 90 年代开发了 DNA 微阵列基因芯片，该技术使研究人员可以同时测定成千上万个基因的表达水平。在现代生物医学中，基于基因表达数据的肿瘤分型诊断已成为 DNA 微阵列的重要应用领域[1-3]。

通常，基于微阵列数据的肿瘤分型诊断需要完成两个任务：基因选择和模式分类[4]。由于上千基因中只有部分基因的表达对于肿瘤诊断是显著作用的，基因选择就是识别出那些对于肿瘤分型表达显著的信息基因，一方面可以提高下一步模式分类的精度，另一方面，为生物学家分析肿瘤的病因提供了可靠依据。模式分类则需要设计一个合适的分类模型，根据已知类别的数据训练出分类决策规则或者构建分类器；进而对未知类别的数据进行预测，也就是生物意义上的肿瘤分型诊断[1,4]。

2000 年 Alizadeh 等[5]利用基因表达数据诊断出病理学尚未发现但临床上却很重要的 DLBCL 子类型。随后几年里，多种传统的机器学习以及模式识别分类方法应用于微阵列数据，从而实现肿瘤分型的分类诊断，如线性判别分析（LDA）、logistic 回归判别法、K 最近邻（K-NN）、支持向量机

* 收稿日期：2008 - 07 - 01
基金项目：国家自然科学基金资助项目（60575004）
作者简介：石玉（1982 年生），女，博士生；E-mail：shiyu@mail2.sysu.edu.cn

（SVM）以及 Tibshirani 等[6]提出的最近收缩质心方法等。Dudoit 等[7]在 Lymphoma、Leukemia 以及 NCI60 三个数据库中对比了 K-NN、LDA、分类树以及 bagging、boosting 等方法的分类性能。Lee 等[8]在 2005 年对此做了扩充，对微阵列数据分析中常用的 21 种方法、7 个数据库、3 个基因选择方法分别做了比较。

高斯混合模型（GMM）可以应用于有监督学习和无监督学习[9]，在统计学习领域应用非常广泛[10]。在基因聚类的应用中，基于高斯混合模型的有监督聚类方法比无监督聚类以及 SVM 方法有更好的表现[11]，本文引入 GMM 来实现肿瘤数据的分类。

对高斯混合模型的估计，通常会面对一个退化问题，导致参数的极大似然估计（MLE）不存在，增加罚函数通常可以有效解决退化问题。Pan[12]提出惩罚类均值的方法，通过对均值做 L_1 惩罚，得到阈值准测，从而在聚类的同时实现变量选择。该方法在基于无监督聚类模型中，得到非常好的结果。本文把这种思想引入到 GMM 的有监督学习中，实现了在分类过程中同时做进一步的信息基因选择，并在理论上给出这一过程的解释。

在传统的分类方法中，分类器的设计通常独立于基因的选择，这使得肿瘤的诊断与信息基因的选择是无关的两个过程。在进行特征变换以后，虽然分类效果有所提高，但使生物学家解释这一生物过程变得困难。相比以往的方法，本文提出的分类模型将肿瘤的分类与基因的进一步选择有机结合在一起，不但使分类精度得到一定的提高，在分类过程中实现基因选择，使其更具有生物意义。在理想的模拟数据中，该方法显示出优良的结果。此外，在五个公共的微阵列数据集中，我们与几个传统的方法做了比较。结果显示本文的方法可以用较少的变量达到更高的识别率，对参数的选择表现稳定，且不受类数的限制。

1 我们的方法

1.1 模型介绍

一次微阵列实验能获得细胞在某一条件下的基因表达数据，包含成千上万个基因在细胞中的相对或绝对丰度，不同条件（细胞周期的不同阶段，药物作用的不同时间，不同肿瘤类型，不同病人等）下的基因表达数据构成一个 $p \times n$ 的矩阵 X，其中 p 代表基因的数目，n 代表条件的个数，通常情况下 $p >> n$。矩阵 X 的每一个元素 x_{ij} 表示第 i 个基因在第 j 个条件下的表达水平值。行向量 $X_i = (x_{i1}, x_{i2}, \cdots, x_{in})$ 代表基因 i 在 n 个条件下的表达水平，称为基因 i 的表达谱，列向量 $X_j = (x_{1j}, x_{2j}\cdots, x_{pj})^T$ 代表某一条件下各基因的表达水平。

假设列向量 $x_j \in \mathbb{R}^p$，也就是第 j 个样本，是有如下混合高斯模型概率密度的随机变量：

$$f(x_i) = \sum_{k=1}^{K} \pi_k f_k(x_i \mid \mu_k, \sum\nolimits_k)$$

其中 π_k 满足 $0 \leqslant \pi_k \leqslant 1$ 和 $\sum_{k=1}^{K} \pi_k = 1$，是第 k 个聚类的混合比例。

$$f_k(x_i \mid \mu_k, \sum\nolimits_k) = \frac{1}{(2\pi)^{P/2} \mid \sum_k \mid^{1/2}} \cdot$$

$$\exp\left(-\frac{1}{2}(x_i - \mu_k)^T \sum\nolimits_k^{-1} (x_i - \mu_k)\right), k = 1, \cdots, K,$$

是第 k 个聚类的概率密度，是均值为 μ_k 方差为 $\sum_k$ 的正态分布。

指示向量 $z_i = (z_{i1}, z_{i2}\cdots, z_{ik})$，其中 $z_{ik} \in \{0, 1\}$，$(k = 1, 2, \cdots, K)$。若 x_i 属于第 k_0 类，也就是相应于第 k_0 个混合成分则 $z_{ik_0} = 1$，其余的 $k = 1, \cdots, k_0 - 1, k_0 + 1, \cdots, K$ 时，$z_{ik} = 0$。对于训练样本，其类标是已知的，因此指示向量也是确定的。对于检验样本，其类标通过最小化经验期望误差确定，即：

$$k_0 = \arg \max_{1 \leqslant k \leqslant K} \pi_k f_k(x_i \mid \mu_k, \sum\nolimits_k)$$

模型的参数 $\theta = \pi_1, \cdots, \pi_{k-1}, \mu_1, \cdots, \mu_k, \sum_1, \cdots, \sum_k$ 通常采用极大似然估计，其似然函数为：

$$l(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} [\log \pi_k + \log f_k(x_i \mid \mu_k, \sum\nolimits_k)]$$

众所周知，如果 $\sum_k$ 奇异，则对数似然函数是无界的，因此极大似然估计不存在。Pan[12]在解决一个无监督聚类问题时提出一个对均值做一个 L_1 惩罚的方法解决了退化问题。同时，通过阈值实现变量选择，从而得到一个稀疏解。其惩罚的似然函数为：

$$l(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} [\log \pi_k + \log f_k(x_i \mid \mu_k, \sum\nolimits_k)] - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} \mid \mu_{kj} \mid$$

本文将这种模型推广到有监督的分类问题中，从而充分利用已知类属的样本信息，找出对其类属贡献最大的变量，估计分类模型，进而实现对新样本的分类。也就是找出信息基因，并进行肿瘤分型诊断。对这一过程，本文在理论上给出进一步的分

析，并对基因选择给出模型解释。

1.2　参数估计

我们用 EM 算法估计模型中的均值，协方差以及混合比例；用留一交叉验证法估计模型中的惩罚参数。本节重点说明协方差与惩罚参数的估计，均值的估计我们将结合模型的解释在第 1.3 节中给出。

文献［13］从几何意义上讨论了高斯判别分析中协方差的 14 种特征分解，并分别讨论了不同特征分解下的参数估计。其中 $\sum = \lambda B$ 的特征分解模型是普遍较好的。又考虑到本文数据的标准化处理，我们选取该对角模型，即：$\sum_k = \sum = \mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_p)$。用 EM 算法求解 MLE，初值的选择常常影响解的效果以及求解的速度。本文充分利用样本的类属信息，参数的初始值分别采用类内样本均值

$$\hat{\mu}_k = \bar{x}_k = \frac{\sum_{i/z_i=k} x_i}{n_k}$$

和全局样本方差

$$\hat{\sum} = \frac{\mathrm{diag}(\sum_{k=1}^{K} n_{1k}\hat{\sum}_k)}{n_1}$$

其中，

$$\hat{\sum}_k = \frac{\sum_{i/z_i=k}(x_i-\bar{x}_k)(x_i-\bar{x}_k)'}{n_{1k}},\quad \forall k = 1,\cdots,K$$

由于样本方差是对角的，其估计方法可参考文献［13］。

作为模型参数的选择，交叉验证法是一种最简单应用最广泛的方法。微阵列数据最大的特点就是维数高，样本少。因此我们选用留一交叉验证法[10]，从而避免因样本较少而引起过高地估计预测误差。此外，不同的惩罚参数可能其预测误差是相同的，在这种情况下，本文自适应选择使识别率达到最高的最大 λ。这样的选择使得用较少的变量达到较高的识别率，从而提高效率并为进一步找到对分类贡献最大的变量提供依据。对模拟数据的实验显示出这样的选择是非常有意义的。

1.3　模型分析与解释

本文的分类方法相比以往的方法，最大的不同在于在分类过程中能自适应地进行变量选择，这一过程使得分类方法与变量选择有机的结合在一起。在分类模型参数估计中，兼顾到变量选择的因素，使模型的估计达到效果最优。下面我们就这一过程的实现给出解释。

在计算模型的参数时，用 EM 算法求解模型的极大似然估计。与传统 EM 算法不同的是，这里的似然函数多了一个惩罚项 $\lambda\sum_{k=1}^{K}\sum_{j=1}^{p}|\mu_{kj}|$，因此在 M 步极大化似然函数过程中，有

$$\frac{\partial l}{\partial \mu_k} = \sum_{i=1}^{n} z_{ik}\sum^{-1}(x_i-\mu_k) - \lambda\,\mathrm{sgn}(\mu_k)$$

文献［12］中，称这一过程为极大化惩罚似然函数（MPLE）。经过一些简单的代数推导，可以得到

$$\mu_k^{(t)} = \mathrm{sgn}(\hat{\mu}_k)\left(|\hat{\mu}_k| - \frac{\lambda}{n_{1k}}\sum^{(t-1)}1\right)_+$$

其中表示每个元素都为 1 的一个 p 维列向量。这一迭代过程，使得解收敛时 μ_k 是稀疏的，即 μ_k 的一些分量为 0。注意到模型中的判别函数可做如下分解：

$$k_0 = \arg\max_{1\le k\le K}\pi_k f_k(x_i\mid\mu_k,\sum_k) =$$

$$\arg\max_{1\le k\le K}\pi_k\exp\left(-\frac{1}{2}(x_i-\mu_k)^{\mathrm{T}}\sum_k^{-1}(x_i-\mu_k)\right) =$$

$$\arg\max_{1\le k\le K}\left\{\log\pi_k - \frac{1}{2}(x_i-\mu_k)^{\mathrm{T}}\sum_k^{-1}(x_i-\mu_k)\right\} =$$

$$\arg\max_{1\le k\le K}\left\{\log\pi_k - \frac{1}{2}\sum_{j=1}^{p}(x_{ji}-\mu_{kj})^2/\sigma_j^2\right\} =$$

$$\arg\max_{1\le k\le K}\left\{\begin{array}{l}\log\pi_k - \frac{1}{2}\sum\limits_{j\mid\mu_{kj}\neq 0}(x_{ji}-\mu_{kj})^2/\sigma_j^2 - \\ \sum\limits_{j\mid\mu_{kj}=0}(x_{ji}-\mu_{kj})^2/\sigma_j^2\end{array}\right\} =$$

$$\arg\max_{1\le k\le K}\left\{\begin{array}{l}\log\pi_k - \frac{1}{2}\sum\limits_{j\mid\mu_{kj}\neq 0}(x_{ji}-\mu_{kj})^2/\sigma_j^2 \\ -\sum\limits_{j\mid\mu_{kj}=0}x_{ji}^2/\sigma_j^2\end{array}\right\}$$

由最后一个等式可以看出：若对任意的 $k(k=1,\cdots,K)$，有 $\mu_{kj}=0$，则第 j 个分量对分类不起作用，即第 j 个基因在这些肿瘤类型中不显著。若对某个 k，有 $\mu_{kj}=0$，则说明第 j 个分量对该类不起作用，即第 j 个基因与第 k 个癌症无关。所以，我们的模型得到稀疏的 μ_k，也即倾向于选择那些与癌症分类有关的基因。

1.4　算法流程

本文的算法主要分为预处理、参数估计、预测分类三个部分。

第一步：预处理

预处理包括数据的标准化和利用 BSS/WSS 准测[7]进行基因的预选择两步。为方便标记，预处理后的数据依然标记为 $X_{p\times n}$。

通过标准化，使每个基因的表达谱平均值为

0，标准差为1。这样可以在同一标准下做比较，消除量纲以及噪音的影响。为了尽可能提取有效信息，传统的分类方法通常先进行特征提取。常用的选择方法有 Wilcoxon 统计量，BSS/WSS 准则，软阈值处理等。文献［8］在不同的分类器中对比了以上方法集。此外经典的 PCA 方法也应用到基因的特征提取中。但 PCA 的实质是进行了特征变换，选出的特征并不是直观的基因表达数据，而是提取的抽象信息。这为生物医学意义上的解释带来了困难。因此本文仍然选用简单、直观的 BSS/WSS 准则进行基因的预选择。关于 BSS/WSS 的详细信息可参考［7］。

第二步：参数估计

本文用极大化惩罚似然函数的方法，估计模型的参数。

随机取出数据的2/3做为训练样本用于估计参数（$i = 1,\cdots,n_1$），剩下1/3做为检验样本做为预测分类（$i = n_1+1,\cdots,n$）。类内均值μ_k的计算参考本文1.3中，协方差 $\sum$ 的计算可参考［13］。类似于文［12］提到的 MPLE，这是一个 EM 算法的迭代求解过程：

（0）初始化类属概率、混合比例、样本均值、以及协方差。为了提高迭代的速度，我们分别采用样本均值、样本方差做为类内均值和协方差的初始估计值。

$$p_{ik}^{(0)} = 1/K, \quad \forall i = n_1+1,\cdots,n; k = 1,\cdots,K$$

$$\pi_k^{(0)} = n_{1k}/n_1, \quad \forall k = 1,\cdots,K$$

$$\mu_k^{(0)}\hat{\mu}_k, \quad \forall k = 1,\cdots,K$$

$$\sum{}^{(0)} = \hat{\sum} = \frac{\mathrm{diag}(\sum_{k=1}^{K} n_{1k}\hat{\sum}_k)}{n_1}.$$

（1）计算均值，协方差。

$$\mu_k^{(t)} = \mathrm{sgn}(\hat{\mu}_k)\left(|\hat{\mu}_k| - \frac{\lambda}{n_{1k}}\sum{}^{(t-1)}1\right)_+$$

$$\forall k = 1,\cdots,K$$

$$\sum{}^{(t)} = \frac{\mathrm{diag}(\sum_{k=1}^{K}\sum_{i/z_i=k}(x_i-\mu_k^{(t)})(x_i-\mu_k^{(t)})'))}{n_1}$$

（2）计算后验类属概率，混合比例。

$$p_{ik}^{(t)} = \frac{\pi_k^{(t-1)} f_k(x_i \mid \mu_k^{(t)}, \sum{}^{(t)})}{\sum_{k'=1}^{K}\pi_{k'}^{(t-1)} f_k(x_i \mid \mu_{k'}^{(t)}, \sum{}^{(t)})}$$

$$\forall i = n_1+1,\cdots,n; k = 1,\cdots,K$$

$$\pi_k^{(t)} = \frac{n_{1k} + \sum_{i=n_1+1}^{n} p_{ik}^{(t)}}{n}, \quad \forall k = 1,\cdots,K$$

（3）检验是否收敛．如果收敛，迭代终止；反之，转（1）。

第三步：预测分类

将检验样本，和第二步得到的模型参数，代入到判别函数中

$$k_0 = \arg\max_{1\le k\le K}\pi_k f_k(x_i \mid \mu_k, \sum{}_k)$$

即可得到每个样本的预测分类。

2 实验结果与分析

2.1 模拟数据

为了验证我们方法的有效性，首先我们构造一个模拟数据做检验。该模拟数据是100个1000维的样本，样本共来源于三类。其中第一类有30个样本，只与40个变量有关，独立同分布产生于$N(2,1)$。第二类也有30个样本，40个显著变量，独立同分布产生于$N(-2,1)$。剩下40个样本属于第3类，也有40个显著变量，独立同分布产生于$N(4,1)$。第一，二类与第三类的显著变量之间无交叉，第一类与第二类有20个共同的显著变量。其余900个变量为与分类无关的白噪声$N(0,1)$。详情参考图1。

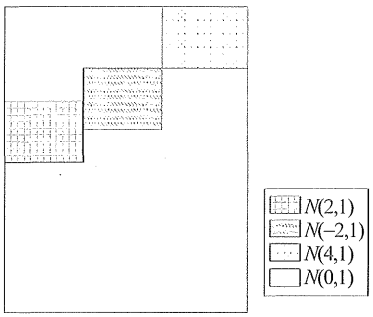

图1 模拟数据

Fig. 1 Simulation data

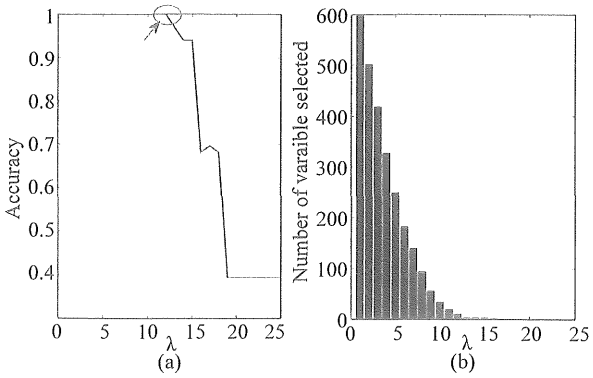

图2 （a）表示不同惩罚参数相应的识别率，（b）表示不同惩罚参数相应的变量个数

Fig. 2 （a）Accuracy under each penalty value，（b）Number of variable selected under each penalty value

在模拟数据中，随机选取其中三分之二作为训练样本，剩下三分之一做检验。做以下两个实验。

实验一：分别预选 100、150、200、400、600 个较显著的变量，用本文的方法做分类，每种情况实验重复 50 次，所有的识别率均达到了 1。这说明我们的方法是非常有效的。对于不同的变量预选择，文中的分类方法在分类过程中均能自动调整惩罚参数，实现变量选择，从而使分类有稳定的表现。惩罚参数的选择是用留一交叉验证在 0,1,2,…,25 中自适应选择。

实验二：实验中，跟踪惩罚参数的选择，观察不同惩罚参数对分类效果和变量选择的影响。图 2 给出变量预选择 600 个时不同惩罚参数下的识别率以及相应的变量个数，模型最终选择参数 $\lambda = 11$，识别率达到 1，选择的变量个数为 10。如图 2（b）中所示，随着 λ 的增大，变量的个数迅速减少，但图 2（a）中识别率变换缓慢。λ 从 0 到 11，图 2（a）识别率均为 1，图 2（b）中选择的变量迅速由 600 变到 10。这说明预选的 600 个变量中，只有部分变量对分类起作用，其他大部分变量都对分类不起作用，或者其中一些是起协同作用的。$\lambda = 0$ 时，相当于经典的高斯混合模型，而本文的算法用较少的变量数可以达到与高斯混合模型相同的性能。λ 从 11 变化到 12，选择变量的个数由 10 变化到 3，识别率降低；说明其中的 7 个变量对分类是有贡献的，被惩罚掉以后信息损失，降低了识别率。但 λ 从 12 变化到 14 识别率又保持稳定，说明这中间惩罚掉的变量与留下的变量起协同作用，有冗余的信息，因此惩罚掉一两个变量，信息并没有减少，因此识别率可以保持稳定。

我们的模型选择如图 2（a）中箭头所指的 λ，它正好是识别率出现的向下的拐点位置，这使得我们用尽可能少的变量，达到尽可能高的识别率。实验的结果显示参数 λ 有效实现对变量个数的控制并保证识别率。

2.2　肿瘤数据

我们在五个公共基因库上探讨我们的方法，并与其他几种常见的方法做比较，用以检验该方法在实际应用中的表现。

2.2.1　*数据介绍*　我们有意选择 5 个不同类型，不同类数的数据集，来探讨本文方法的性能。数据涵盖了正常组织与肿瘤组织的，以及肿瘤在不同子类型之间的分类。下面是我们分别采用的 2，3，4，5，6 类问题的几个数据集：

结肠癌（Colon）：这个数据集源于对结肠癌研究的寡核苷酸微阵列，其中包含 40 个肿瘤结肠，20 个正常结肠的组织样本；每一个包含 6500 多个基因的表达，我们采用文献［14］中选取的 2000 个基因的表达。

白血病（Leukemia）：这个数据集包含了 3571 个基因在 72 例白血病人中的表达数据，可以分为 B 细胞急性淋巴白血病（B-ALL），T 细胞急性淋巴白血病（T-ALL）和急性骨髓白血病（AML）三种子类型。其中有 38 例为 B-ALL，9 例为 T-ALL，其余 25 例为 AML[15]。

小圆形蓝色细胞瘤（Small，Round Blue Cell Tumors，SRBCT）：SRBCT[15]是一个关于儿童小圆形蓝色细胞瘤的数据集包括含有 2308 个基因表达的 83 例病变样本，分布在 4 个子类型中。

表 1　数据集的结构

Table 1　Statistics of data sets

数据集	样本个数及构成	基因个数	类个数
Colon	22 + 40 = 62	2 000	2
Leukemia	38 + 9 + 25 = 72	3 571	3
SRBCT	23 + 8 + 12 + 20 = 63	2 308	4
Brain	10 + 10 + 10 + 4 + 8 = 42	5 597	5
ALL	15 + 27 + 64 + 20 + 43 + 79 = 248	12 625	6

脑癌（Brain）：这个数据集[16]中引入，包含了 5579 个基因在 42 例脑癌病人中的表达数据。数据包含 5 个类，其中 10 个成神经管细胞瘤（medulloblastoma），10 个恶性神经胶质瘤（malignant glioma），10 个 AT/RT，8 个 PNET，以及 4 个正常的小脑。

急性淋巴白血病（Acute Lymphoblastic Leukemia，ALL）：这个数据集包含 6 个急性淋巴白血病的子类型[17]。

2.2.2　*实验结果*

实验一：所有数据均进行了标准化处理，每个数据集我们随机的分为 2/3，1/3 两部分。其中 2/3 的部分作为训练集估计模型参数，剩下 1/3 作为检验集对我们的分类模型进行检验。这个过程重复 50 次，下文的识别率取其平均值，同时给出标准差展示方法的稳定性。在同一条件下，我们进行多种方法的比较。

表 2 和图 3 是基因预选择 200 个的情况个，各种方法的识别率比较，本文的方法简写为 pnlGMM。Dudoit 等[7]认为基因预选 200 个对大部分分类器的性能影响都不大。对比几个现有的方法，

我们发现与文［8］中研究结果一致。K-NN 对比复杂的方法，有更好的表现；PCA + LDA、DLDA 和 Diagonal LDA 三个分类器非常不稳定，其中 DLDA 在类数较多的 Brain、ALL 数据集中有着较高的识别率。相比之下，我们的方法分类效果最好。在 Brain、SRBCT 和 Colon 数据集上，识别率高于 K-NN。可以看出，本文方法虽然在 Leukemia 中的表现稍差一些，但仍然保持着较高的识别率。因此整体上来讲，本文的方法分类效果精确，对不同数据库表现稳定。

表 2 几个不同方法之间的比较

Table 2 Accuracy comparison with other methods

方法/数据集	Brain	SRBCT	Colon	Leukemia	ALL
pnlGMM	94.14 ± 5.14	99.05 ± 1.92	88.10 ± 7.19	93.50 ± 4.85	97.54 ± 1.54
DLDA	92.64 ± 7.56	88.76 ± 7.12	83.93 ± 5.94	77.29 ± 8.44	96.80 ± 1.82
PCA + LDA	93.57 ± 6.46	75.29 ± 9.61	86.31 ± 5.64	61.98 ± 10.59	89.87 ± 3.71
Ye's ULDA	93.86 ± 6.26	98.62 ± 2.27	86.61 ± 5.87	97.40 ± 2.92	81.43 ± 4.29
K-NN	92.71 ± 6.50	98.57 ± 2.49	85.83 ± 6.11	96.25 ± 3.15	95.05 ± 2.30
Diagonal LDA	93.50 ± 6.43	35.21 ± 8.35	87.56 ± 5.22	85.16 ± 6.35	97.65 ± 1.61
Ye's RDA	94.02 ± 5.62	98.33 ± 2.74	87.20 ± 5.90	95.26 ± 4.38	97.61 ± 1.50
GLDA	93.57 ± 6.18	98.57 ± 2.32	86.61 ± 5.87	97.40 ± 2.92	81.17 ± 4.46

实验二：该实验考察本文方法在不同基因预选择情况下的表现。在实验一的数据条件下，分别用 BSS/WSS 准则预选出 100、150、200、400、600 个基因，再用本文的分类方法实现。表3 和图 4 给出实验结果。从图表中可以看出我们的方法在不同的基因预选情况下，识别率相差不大。这说明我们的算法对于不同的参数选择，结果都相当稳定。相比之下，预选 150 ~ 200 个基因，可以在每一个数据库中达到稳定较高的识别率，这与文［8］中关于基因选择的研究结果一致。此外，基因预选个数多的时候，维数增高，会带来计算复杂度增大。表 4 是在数据集 SRBCT 上，在相应基因个数下，配置为 Inter（R）Xeon（TM）CPU 3.06GHz，RAM 2.50GB的计算机运行一次实验花费的时间。因此基因预选 150 ~ 200 个，是一个不错的选择。

表 3 本文方法在不同基因预选择数目下的识别率

Table 3 Accuracy comparison between different numbers of gene preselected

预选个数/数据集	Brain	SRBCT	Colon	Leukemia	ALL
100	91.29 ± 5.82	98.29 ± 2.31	88.19 ± 6.03	92.58 ± 4.86	96.14 ± 2.01
150	94.00 ± 5.47	99.14 ± 1.85	86.57 ± 5.98	93.08 ± 4.66	97.69 ± 1.63
200	94.14 ± 5.14	99.05 ± 1.92	86.95 ± 5.68	93.50 ± 4.85	97.54 ± 1.54
400	92.00 ± 5.71	98.76 ± 2.32	86.95 ± 5.68	93.25 ± 4.60	97.59 ± 1.61
600	92.86 ± 5.77	98.76 ± 2.32	87.14 ± 5.79	95.08 ± 4.01	97.49 ± 1.74

表 4 本方法在不同基因预选择情况下需要的时间

Table 4 Time expended by proposed method on different number of gene preselected

预选基因个数	100	150	200	400	600
运行时间/min	0.046 617	0.117 18	0.167 97	1.557 6	5.059 4

3 结 语

对微阵列基因表达数据的分析，为现代医学进行肿瘤诊断与治疗提供了准确依据。微阵列数据的高通性，使得一次实验即可同时测定成千的基因的表达。因此在分类的过程中，兼顾变量的选择是非常必要的。本文提出一种带惩罚的高斯混合分类器，使分类器的设计与变量的选择有机结合在一起，从而用较少的变量，达到更好的分类效果。此外，该方法不受类的数目限制，在两类和多类问题上，均有理想的识别效果。对于变量的预选择个数，本文的方法并不敏感，对不同的变量个数，均

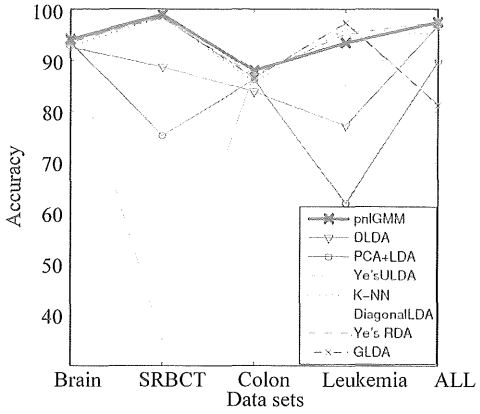

图 3　8 个方法在 6 个数据库上识别率的对比

Fig. 3　Accuracy comparison between 8 methods on 6 datasets

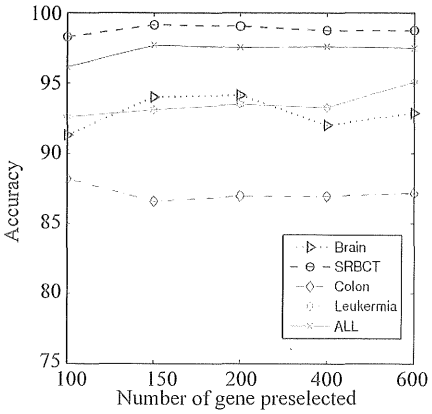

图 4　本文方法在不同基因预选择数下的识别率

Fig. 4　Accuracies of proposed method on different number of gene preselected

可自适应调整，进行变量的再选择。该方法同样适用于其他高维多类问题，对于需要对变量做出解释的应用领域，本文的方法值得推荐。对于不需要做变量解释的，则变量预选择可以采用 PCA，PLS 等特征变换的方法代替 BSS/WSS 准则。在模拟实验和五个公共基因数据集上的应用显示了本文的方法具有精确的识别效果和稳定性。

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] SCHENA M. Microarray analysis [M]. NJ: Wiley-Liss Hoboken, 2002.

[2] BALDI P, HATFIELD G W. DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling [M]. Cambridge University Press, 2002.

[3] SCHMIDT U, BEGLEY C G. Cancer diagnosis and microarrays [J]. International Journal of Biochemistry and Cell Biology, 2003, 35:119 - 124.

[4] LIN T C, LIU R S, et al. Pattern classification in DNA microarray data of multiple tumor types [J]. Pattern Recognition, 2006, 39:2426 - 2438.

[5] ALIZADEH A A, EISEN M B, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling [J]. Nature, 2000, 403:503 - 511.

[6] TIBSHIRANI R, HASTIE T, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression [J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99:6567 - 6572.

[7] 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.

[8] LEE J W, LEE J B. An extensive comparison of recent classification tools applied to microarray data [J]. Computational Statistics and Data Analysis, 2005, 48:869 - 885.

[9] CELEUX G, GOVAERT G. Gaussian parsimonious clustering models [J]. Pattern Recognition, 1995, 28:781 - 793.

[10] HASTIE T, TIBSHIRANI R, FRIEDMAN J. The Elements of Statistical Learning [M]. New York: Springer-Verlag, 2001.

[11] QU Y, XU S Z. Supervised cluster analysis for microarray data based on multivariate Gaussian mixture [J]. Bioinformatics, 2004, 20:1905 - 1913.

[12] PAN W, SHEN X T. Penalized model-based clustering with application to variable selection [J]. Journal of Machine Learning Research, 2007, 8:1145 - 1164.

[13] BENSMAIL H, CELEUX G. Regularized Gaussian discriminant analysis through eigenvalue decomposition [J]. Journal of the American Statistical Association, 1996, 91:1743 - 1748.

[14] ALON U, BARKAI N, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays [J]. Proceedings of the National Academy of Sciences, 1999, 96:6745 - 6750.

[15] GOLUB T R, SLONIM D K, TAMAYO P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring [J]. Science, 1999, 286:531 - 537.

[16] POMEROY S L, TAMAYO P, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression [J]. Nature, 2002, 24:436 - 442.

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