秩和基因选取方法及其在肿瘤诊断中的应用

邓 林① 马尽文①* 裴 健②

(①北京大学数学科学学院信息科学系，数学与应用数学重点实验室，北京 100871; ②Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260-2000, USA. * 联系人, E-mail: jwma@math.pku.edu.cn)

摘要 根据基因表达谱进行肿瘤诊断是当今生物信息学领域中的一个重要研究方向，其中最主要的问题是肿瘤相关基因的选取. 根据统计学中的秩和检验方法提出了秩和基因选取方法. 并利用支持向量机(SVM)对相关基因表达谱数据进行训练建立肿瘤诊断模型. 实验表明这种方法与模型可使得在结肠数据和白血病数据上的诊断正确率分别达到 96.2%和 100%.

关键词 基因表达谱 秩和方法 支持向量机 肿瘤诊断 基因选取

随着 DNA 微阵列技术的快速发展，基因表达谱数据的获得已变得越来越快捷和可靠. 这些生物数据为人体组织的健康状况和病症分析与识别提供了重要依据. 如何从基因表达谱数据中分析出有价值的生物学信息已成为当今生物信息学研究的主要课题[1~8].

基因表达谱数据一般表示成一个基因表达矩阵 $W=(w_{ij})_{n\times m}$，如图 1 所示. 其中第 i 行对应于第 i 个基因，第 j 列对应于第 j 个样本(病例)，元素 w_{ij} 则表示第 j 个样本关于第 i 个基因的 mRNA 表达水平. 通过对基因表达谱数据的分析，生物学家们能够获得大量有价值的生物学信息. 近几年来，基于基因表达谱的分析研究已被广泛应用于肿瘤分类与诊断及其基因生物功能的确定等方面. 其常用的分析方法包括聚类、分类和主成分分析等. 特别地，基于基因表达谱对肿瘤进行分类与诊断已成为其中一个重要研究方向[1~6]. 1999 年，Golub 等[1]首先采用邻域分析方法对白血病进行分类，并在此过程中采用了一种 t 统计量的简化形式作为辨识性度量选取了 50 个最相关基因构建分类器. 同年，Alon 等[2]对结肠的基因表达谱做了聚类分析，得到了一些表达谱与肿瘤的对应关系，其中同样使用了 t 统计量方法进行相关基因选取. 2000 年，Brown 等[3]将几种常用分类方法应用到基于基因表达谱的肿瘤分类，并对分类效果进行了比较，发现采用支持向量机(SVM)效果最好. 这一结论也被 Dudoit 等[4], Furey 等[5]和 Guyon 等[6]的研究结果所进一步证实.

这些研究表明，基于基因表达谱的肿瘤分类与诊断是可行和可靠的. 然而，如果不对基因表达谱数据进行预处理，便直接投入到分类方法当中，所得到的结果往往很不理想. 主要表现在肿瘤分类方法的推广能力不足，即根据训练样本集所得到的分类规则在检验样本集上表现出较低的正确率，即使采用推广能力很好的 SVM 也是如此. 我们认为其主要症结在于没有很好的剔除基因表达谱中的噪声. 实际中，某类肿瘤的出现可能仅仅与某些基因的表达水平的变化有关. 若笼统地用全部基因表达水平来进行分类，不仅会因数据维数的巨大而难于进行，而且众多无关数据将便成噪声大大地干扰分类的结果. 为此人们已经提出了一些相关基因选取的方法[1,2,5~8]. 其中现阶段应用最广泛的是 t 统计量方法及其变形. 而 t 统计量方法的统计学依据是 t 检验. 我们知道 t 检验是一种参数检验方法，假设样本总体服从正态分布. 因此 t 统计量方法及其变形都是以基因表达谱服从正态分布的假设为依据，而实际发现这一假设常常并不成立(见下节分析).

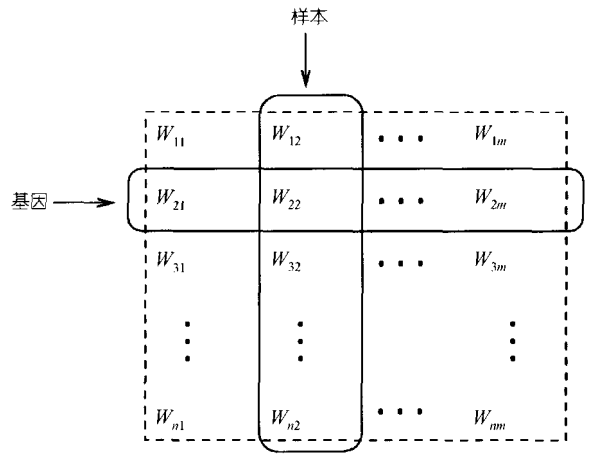

图 1 基因表达谱矩阵

为了避免正态假设，我们依据非参数统计中的秩和检验理论提出了秩和相关基因选取方法. 然后，采用 SVM 建立肿瘤诊断模型，并根据简化后的训练样本数据(相关基因表达谱)进行有监督的学习，最后在检验样本数据集上进行检验. 通过对两类肿瘤基因表达谱数据的训练和检验，我们发现这种秩和基因选取方法可以使得 SVM 分类器获得很高的推广能力.

下文中，我们将在第 1 节中首先对于相关基因的统计方法进行理论分析，然后提出了秩和基因选取方法，并进一步引出了 SVM 作为肿瘤诊断模型. 在第 2 节中，我们首先给出了采用秩和方法进行基因选取并应用 SVM 进行肿瘤诊断的一些实验结果，然后再与 t 统计量方法的结果进行了比较. 最后在第 3 节给出结论.

1 秩和基因选取方法与肿瘤分类模型

1.1 相关基因选取的统计分析研究

人们很早便开始了肿瘤相关基因的识别研究，但基本上是根据生物特性进行的. 随着 DNA 微阵列技术的发展和基于基因表达谱的肿瘤分类方法的研究，人们提出了一些基于统计分析的肿瘤相关基因选取方法. 这些方法通过引入基因对肿瘤的辨识性度量，选取出对肿瘤辨识性较大的基因.

特别地，t 统计量及其变形是现今最常用的肿瘤辨识性度量. t 统计量的表达式为 $T=\dfrac{m_{i,+}-m_{i,-}}{S_w\sqrt{\frac{1}{n_+}+\frac{1}{n_-}}}$，其中 $S_w^2=\dfrac{(n_+-1)s_{i,+}^2+(n_--1)s_{i,-}^2}{n_++n_--2}$；式中的 $m_{i,+}$，$m_{i,-}$，$s_{i,+}$ 和 $s_{i,-}$ 分别为第 i 个基因在正、负样本中表达水平的均值和标准差，n_+ 和 n_- 分别是正负样本的个数. 事实上，t 统计量在统计学的二元 t 检验中可以用于度量两个正态总体的分布差异大小. 因此 T 的绝对值越大，意味着该基因的表达水平的在正负样本中变化越显著，该基因与样本的肿瘤因素的相关性也就越大. 换句话说该基因对肿瘤的辨识性越强. t 统计量方法[2]基于这一统计直观选取出 T 的绝对值最大的 K 个基因为相关基因.

由于 t 统计量的表达式较复杂，人们也提出了一些简化的表达式. 例如 Golub 等[1]采用的辨识性度量为 $W_i=\dfrac{m_{i,+}-m_{i,-}}{s_{i,+}+s_{i,-}}$，并要求所选取基因的 W_i 取正值和负值的个数相同. 另外，Furey 等[5]采用了 W_i 的绝对值，Pavlidis 等[7]采用了 W_i 二次值作为辨识性度量. t 统计量及其变形的表达式本质上是一致的，其分子都是正负样本基因表达水平的均值的差，分母都是正负样本基因表达水平的方差的函数，用于正规化辨识性度量的表达式.

此外，Ding[8]提出了 t 统计量的一般化形式 F 统计量作为辨识性度量，可以处理肿瘤类数多于 2 的相关基因选取问题. 假定某基因的表达水平为 $g=(g_1,g_2,\cdots,g_n)$，肿瘤类别数为 K，则 F 统计量可表达为 $F=\left[\sum_k n_k(\bar{g}_k-\bar{g})^2/(K-1)\right]\Big/s^2$，，其中 $\bar{g}$ 和 $\bar{g}_k$ 分别是该基因在全体样本和第 k 类样本中的平均表达水平，n_k 和 s_k 表示第 k 类的样本数和方差，$s^2=\left[\sum_k(n_k-1)s_k{}^2\right]\Big/(n-K)$. 当 $K=2$ 时，$F=t^2$，$t=\sqrt{\dfrac{n_1n_2}{n_1+n_2}}\dfrac{\bar{g}_1-\bar{g}_2}{s}$，F 统计量退化为 t 统计量.

如前所述，t 统计量方法及其变形都是以 t 检验作为其统计依据的. 然而 t 检验是一种参数检验方法，以样本服从正态总体的假设为前提. 那么在基因表达谱不服从正态分布的情况下，使用 t 统计量方法选取相关基因能否得到好的分类结果呢? 从以前的研究成果和我们的实验(见实验部分)可以看出：一般来说，即使样本不完全满足正态条件，利用 t 统计量方法进行相关基因选取依然是有效的. 即利用 t 统计量方法进行基因选取的分类器，比起不做基因选取的分类器在正确率上能有较大提高. 因此，即使正态假设不完全满足，t 统计量仍能在一定程度上能反映出基因表达水平对肿瘤因素的辨识性.

然而从理论上讲，t 统计量在正态假设不满足的条件下对基因的辨识性度量则是不精确的，缺乏牢靠统计依据的. 这时候采用 t 统计量方法会产生两个问题：第一，利用 t 统计量对基因的排序与真实按照基因对肿瘤辨识性大小的排序可能出现不一致，即基因错位. 例如，假设有两个基因 A 和 B，A 的表达水平服从正态分布，B 的表达水平服从某一区间内的均匀分布，A 的 t 统计量的值大于 B 的 t 统计量的值. 因此，A 被排在了 B 的前面. 然而问题的关键在于，真实反映 B 基因对肿瘤因素辨识性大小的 p 值(p-value，或称临界值)[9]应该按照均匀分布而非正态分布的条

表 1　在 0.05 显著性水平下, 3 个数据集的正态性检验结果

	结肠正常样本	结肠肿瘤样本	乳腺正常样本	乳腺肿瘤样本	急性淋巴白血病	急性髓性白血病
总基因数	2000	2000	5776	5776	7129	7129
零假设被否定的基因数	730	1483	2250	2474	4542	2558

件计算. 因此按照 A, B 两基因真实的 p 值排序, B 基因就有可能排在 A 基因的前面. 这就是 t 统计量在非正态条件下造成了基因的错位. 第二, 当基因表达谱不服从正态分布时, 计算出的 t 统计量也不再服从 t 分布, 因此无法通过查统计表或利用统计软件计算该基因真实的 p 值. 得不到 p 值就不能利用显著性水平阈值确定合适的相关基因数目, 这将给用户确定合适的相关基因数目带来不便. 因此从理论上说, t 统计量方法只有在正态性条件满足的情况下才能表现得最好. 然而, 正态性假设无疑限制了 t 统计量方法的应用范围.

为了验证实际问题中基因表达谱是否服从正态分布, 我们在结肠数据 [1)]、白血病数据 [2)]和乳腺数据 [3)]上利用峰度和偏度[9]做了正态性检验. 零假设是样本服从正态分布, 所取的显著性水平为 0.05, 即犯第一类错误的概率小于 5%. 实验结果见表 1. 从表 1 看出, 约半数基因的正态零假设被否定了. 即使是零假设未被否定的基因也不能认为其肯定服从正态分布, 因此显然不能认为上述 3 个数据集的基因表达谱服从正态假设. 由于上述 3 组数据都是非常典型的基因表达谱数据, 因此肿瘤分类问题的基因表达谱不服从正态分布的情况应该是普遍的. 在正态假设不满足的情况下, 需要一种更有统计依据的方法替代 t 统计量方法进行基因选取. 本文提出的秩和方法正是这样一种方法.

1.2　秩和基因选取方法

为了避免正态假设, 我们将非参数统计中的秩和检验理论[10]应用于相关基因的选取, 建立了基因选取的秩和方法. 非参数统计方法的优点是具有样本分布的无关性, 即不需要假设样本分布的类型. 其中秩和检验是一种常用的、检验效率很高的非参数检验方法. 非参数统计的理论证明, 在样本总体不满足正态条件的情况下, 秩和检验的 Pitman 渐进效率远高于 t 检验[11]. 对于相关基因选取问题而言, 检验效率高意味着在相同样本量之下, 由秩和方法得到相关基因更能反映出对肿瘤因素的辨识性.

秩和方法基本思想上是对每个基因的表达水平做秩和检验, 判定肿瘤和正常样本的分布间是否存在显著差异. 如果是, 表明肿瘤因素对该基因的表达水平有着显著影响. 因此该基因与肿瘤关系紧密, 将其选取出来. 与 t 统计量方法直接利用基因表达水平的值相比, 秩和方法先通过对表达水平排序得到“秩”(即序列中的位置), 再计算秩和统计量进行分析. 我们知道, 实际的基因表达谱数据中含有大量的噪声和奇异值. 这些噪声和奇异值会极大的影响 t 统计量的值, 但却不会对利用“秩”的秩和统计量的值造成很大影响, 可以说“秩”平滑了噪声. 因此秩和方法比 t 统计量方法更适合于噪声较多的基因表达谱数据.

根据不同的样本类别数, 秩和检验可以分为 Wilcoxon 秩和检验和 Kruskal-Wallis 秩和检验. 前者用于解决两肿瘤类别(或肿瘤与正常)的相关基因选取问题, 后者适用于多肿瘤类别的基因选取. 本文着重于两肿瘤类别的基因选取问题, 以下简述 Wilcoxon 秩和检验用于基因选取的基本步骤:

(1) 建立假设

一般建立如下假设, H_0: 肿瘤和正常样本(或两类肿瘤)的总体分布相同; H_1: 两类别的总体分布不同; 检验的显著性水平为 a .

(2) 混合编秩

将一个基因在两类样本中的表达水平混合, 从小到大排序, 排序号称为“秩”, 相同数值的用平均秩表示. 例如, 有 k 个相同的表达水平分别排在 $n+1$ 到 $n+k$ 位, 则每一个表达水平的秩都用 $n+\frac{k+1}{2}$ 表示.

(3) 计算秩和统计量

取出样本较少的一类, 其样本量用 n_1 表示, 计算该类的秩和 W. 如果 H_0 假设成立, W 的期望值是

1) 结肠数据, 40 个肿瘤样本, 22 个正常样本, 2000 个基因. 来自 http://www.molbio.princetion.edu/colondata

2) 白血病数据, 47 个急性淋巴白血病样本(ALL), 25 个急性髓性白血病样本(AML), 7129 个基因. 来自 http://www.genome.wi.mit.edu/MPR/data_set_ALL_AML.html

3) 乳腺数据, 13 个正常样本, 14 个肿瘤样本, 5776 个基因. 来自 http://genome-www.stanford.edu/sbcmp

$\frac{(n_1+n_2+1)}{2}n_1$，其中 n_2 为另一类的样本量. 因此，W 与 $\frac{(n_1+n_2+1)}{2}n_1$ 相差越大，H_0 假设成立的可能性就越小. 当 W 超出临界范围时就否定 H_0 假设.

(4) 检验

得到了秩和统计量 W 的值后，可以通过查秩和分布表或者利用工具软件如 SAS、Matlab 等计算出 p 值(临界值). 然后与用户设定的显著性水平 a 比较，若 p 值小于显著性水平 a 就将该基因选取出来.

对于多肿瘤类别的基因选取问题应采用 Kruskal-Wallis 秩和检验，它可以替代 Ding 提出的 F 统计量方法[8]. Kruskal-Wallis 秩和检验的基本步骤与上面步骤类似，本文不再赘述.

1.3 肿瘤分类方法

为了检验秩和基因选取方法的有效性，我们结合支持向量机(SVM)[12]建立肿瘤诊断模型. 这里我们简要回顾一下 SVM 的基本思想.

图 2 是二维情况下 SVM 的示意图. 在图中，实心点和空心点代表两类样本，H 为分类线，H_1、H_2 分别为过各类中离 H 最近的样本且平行于 H 的直线，它们间的距离称为分类间隔(margin). 最优分类线是指能正确分开两类样本，而且使分类间隔最大的分类线. H 为最优分类线时，H_1、H_2 上的训练样本点称为支持向量. 构建 SVM 就是一个解最优分类面的问题.

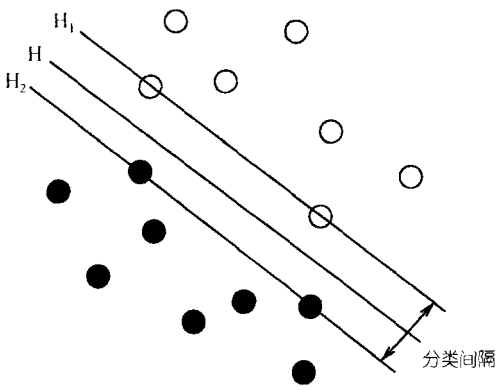

图 2 支持向量机示意图

假设一般分类面方程为 $wx+b=0$. 将其归一化，使得对线性可分样本集：$\{(x_i,y_i),i=1,2,\cdots,N,x_i\in R^d,y_i\in\{\pm1\}\}$，满足约束条件：$y_i(wx_i+b)-1\geq 0,i=1,2,\cdots,N$. 此时分类间隔等于 $2/\|w\|$，使分类间隔最大等价于使 $\|w\|$ 最小. 使 $f(w)=\frac{1}{2}\|w\|^2$ 最小且满足上述约束条件的分类面就是最优分类面.

利用 Lagrange 乘子法可以将上述求解最优分类面问题转化为其对偶问题，即在约束条件 $\sum_{i=1}^{N}y_ia_i=0,a_i\geq 0,i=1,2,\cdots,N$ 下，求优化目标函数 $Q(a)=\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i,j=1}^{N}a_ia_jy_iy_j(x_ix_j)$ 的最大值. 这是一个不等式约束下的二次优化问题，存在惟一解. 可解出最优分类面为 $f(x)=\mathrm{Sgn}\left(\sum_{i=1}^{N}a_i^*y_i(x_ix)+b^*\right)$，容易证明，此解中只有一小部分 a_i^* 不为零，其对应的样本就是支持向量. b^*是分类面的阈值，可以由支持向量得到.

对于线性不可分情况，可以通过非线性变换将问题转化成高维特征空间中的线性问题，在特征空间求解最优分类面. 注意到，在上面的对偶问题中，无论是优化目标函数还是分类函数都只涉及样本内积. 因此，只需用 $K(x,x')$ 代替原先的内积，即相当于将原问题空间变换到新的特征空间. 此时优化的目标函数变为：$Q(a)=\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i,j=1}^{N}a_ia_jy_iy_jK(x_i,x_j)$，而相应的分类函数也变为：$f(x)=\mathrm{Sgn}\left(\sum_{i=1}^{N}a_i*y_iK(x_i,x)+b^*\right)$，其他约束条件不变.

针对不同类型的数据，SVM 可以使用不同的核函数(内积形式). 目前比较常用的有：多项式核函数 $K(x,x_i)=[(xx_i)+1]^q$；径向基(Gauss)核函数 $K(x,x_i)=e^{-\frac{|x-x_i|^2}{s^2}}$ 和 Sigmoid 核函数 $K(x,x_i)=\tanh(v(xx_i)+c)$.

对于肿瘤诊断问题，我们将利用由秩和方法选取出来的相关基因表达谱数据训练 SVM，得到最优分类面. 并依此作为预测未知样本的肿瘤诊断模型.

2 实验结果

2.1 秩和方法的实验结果

为了检验利用秩和方法进行基因选取是否有效，我们在结肠数据和白血病数据上利用秩和方法结合 3

表 2　不同显著性水平 a 取得的相关基因数目表

	总基因数	a = 0.1	a = 0.05	a = 0.01	a = 0.001
结肠数据	2000	210	109	34	8
白血病数据	7129	1837	1425	844	398

种核函数的 SVM 做了肿瘤诊断的实验. 实验程序是用 Matlab6.5 结合 SVM^{light}[13]工具包实现的, 其中 SVM^{light} 用于构建 SVM, 其可在 http://svmlight.joachims.org 上免费下载. 我们在一些常用的显著性水平下, 利用秩和方法对上述两个数据集做了相关基因选取, 结果见表 2.

然后在选取出的相关基因集合上以及未作基因选取的原数据上, 我们使用了线性 SVM、三次多项式 SVM 和径向基 SVM 分别进行了学习和预测. 为了使实验结果更可靠, 我们采用 4-fold cross validation 的方式进行实验. 具体来说, 对于结肠数据, 我们随机将 40 个肿瘤样本和 22 个正常样本分成 4 组, 每组有 10 个肿瘤样本和 5 或 6 个正常样本. 类似地, 我们将白血病数据的 47 个 ALL 和 25 个 AML 样本也随机分成 4 组, 每组有 18 个样本—— 11 或 12 个 ALL 样本加上 5 或 6 个 AML 样本. 然后我们分别做了 4 组实验, 每次将其中 1 组样本用作测试集, 另外 3 组样本用作训练集, 最后计算平均正确率. 在实验中, 为了便于处理, 我们在构建 SVM 之前, 先对原始基因表达谱做了正规化, 使其均值为 0, 方差为 1. 实验结果见表 3 和表 4.

由以上结果可以看出, 利用秩和方法进行基因选取是非常有效的. 在结肠数据上, 利用基因选取后得到的最好正确率是 98.3%, 即在 4-fold cross validation 中只有一个预测错误的样例. 而在白血病数据上, 利用基因选取后甚至取得了 100%的正确率. 平均来说, 在显著性水平 0.01 下利用秩和方法后的正确率在结肠数据上达到 96.2%, 在白血病数据上达到了 100%. 利用秩和方法进行相关基因选取能使肿瘤预测的正确率大幅度提高. 这样的正确率已达到了实际应用的要求.

同时我们看到, 显著性水平的选取是获得较高正确率的关键. 显著性水平取得过高, 一些与肿瘤因素相关性不显著的基因会被选取出来, 成为干扰分类器的噪声, 降低分类正确率; 而显著性水平取得过低, 又会过虑掉一些有用的相关基因, 丢失了信息同样达不到高的正确率. 从以上两组实验来看, 取 0.01 的显著性水平能够使结肠数据和白血病数据获得最好的分类结果. 因此 0.01 的显著性水平对于其它基因表达谱数据也是有一定指导意义的.

2.2　秩和方法与 t 统计量方法的比较

本文前面已经在理论上阐述了, 在正态假设不满足的情况下, 秩和方法比 t 统计量方法更准确更可靠. 这里我们进一步在实验上将秩和方法与 t 统计量方法进行比较. 为了便于对比, 我们让两种方法选取的相关基因数目相同, 所用的实验设定也都与上一小节相同, 不在赘述. 表 5 列出了对于 4-fold cross validation 和 3 种不同类型 SVM 综合的平均分类正确率.

可以看出, 对于不同的相关基因数目秩和方法始终优于 t 统计量方法, 这与我们前面的理论分析是一致的. 除去结肠数据在 0.001 显著性水平下只取得 8 个基因的情况(可能由于基因太少, 不利于比较), 秩和方法的正确率一般高于 t 统计量方法 1%~5%. 考虑到 t 统计量方法的正确率也接近或超过了 90%, 这

表 3　结肠数据结果(a 为显著性水平)

	原数据(2000)	a = 0.1(210)	a = 0.05(109)	a = 0.01(34)	a = 0.001(8)
线性 SVM	56.4%	90.3%	90.3%	95.1%	88.8%
三次 SVM	31.3%	61.3%	90.3%	95.1%	88.8%
径向基 SVM	45.3%	87.2%	93.6%	98.3%	88.8%
平均	44.3%	79.6%	91.4%	96.2%	88.8%

表 4　白血病数据结果(a 为显著性水平)

	原数据(7129)	a = 0.1 (1837)	a = 0.05 (1425)	a = 0.01 (844)	a = 0.001 (398)
线性 SVM	73.6%	94.4%	94.4%	100%	100%
三次 SVM	55.6%	91.7%	91.7%	100%	95.8%
径向基 SVM	52.8%	94.4%	100%	100%	100%
平均	60.7%	93.5%	95.4%	100%	98.6%

表 5 秩和方法与 t 统计量方法的比较

数据集	基因选取方法	210 个基因	109 个基因	34 个基因	8 个基因
结肠数据	秩和方法	79.6%	91.4%	96.2%	88.8%
	t 统计量方法	75.9%	90.3%	93.6%	88.8%
数据集	基因选取方法	1837 个基因	1425 个基因	844 个基因	398 个基因
白血病数据	秩和方法	93.5%	95.4%	100%	98.6%
	t 统计量方法	92.6%	91.7%	95.4%	94.4%

一优势还是十分显著的. 前面的正态性实验表明了结肠和白血病这两个数据集的基因表达谱并不服从正态分布, 因此实验结果验证了秩和方法在这两个数据集上优于 t 统计量方法的理论分析.

3 总结

本文对于基于基因表达谱的肿瘤诊断问题进行了研究, 而肿瘤诊断问题的关键在于相关基因的选取. 本文针对传统的相关基因选取方法依赖正态假设的不足, 提出了秩和基因选取方法. 秩和方法克服了需要正态条件的缺陷, 因此适用于任意的基因表达谱数据. 另一方面, 秩和方法是通过统计假设检验的显著水平来确定出与肿瘤相关基因的个数. 这使得相关基因个数的确定更为科学和合理. 而以往多采用经验的方法来确定相关基因的个数. 进一步, 我们结合支持向量机(SVM)在选取的基因表达谱上进行训练, 建立肿瘤诊断模型. 实验表明, 秩和基因选取方法以及结合 SVM 的肿瘤诊断模型是有效的. 该诊断模型在结肠数据和白血病数据上分别达到了 96.2%和 100%的正确率. 实验也证明了秩和方法是优于传统的 t 统计量方法的. 这些结果表明秩和方法及其结合 SVM 的肿瘤诊断模型是能够应用于实践中.

致谢 本工作为国家自然科学基金(批准号: 60071004)资助项目.

参 考 文 献

1 Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 1999, 286: 531~537

2 Alon U, Barkai N, Notterman D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Nat' l Acad Sci USA, 1999, 96: 6745~6750

3 Brown M P S, Grundy W N, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Nat' l Acad Sci, 2000, 97(1): 262~267

4 Dudoit S, Fridyand J, Speed T P. Comparison of discrimination methods for the classification of tumor using gene expression data. Journal of American Statistical Association, 2002, 97(457): 77~87

5 Furey T, Cristianini N, Duffy N, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 2000, 16(10): 909~914

6 Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machine. Machine Learning, 2002, 46(1/3): 389~422

7 Pavlidis P, Weston J, Cai J, et al. Gene functional Analysis from heterogeneous data. Proc Fifth Int. Conf. on Computational Molecular Biology. New York: ACM Press, 2001. 249~255

8 Ding H Q. Analysis of gene expression profiles: class discovery and leaf ordering. In: Proc RECOMB, 2002. 127~136

9 Goulden C H. Methods of Statistical Analysis. (2nd edition). New York: John Wiley & Sons, 1956

10 Hettmansperger T P. Statistical Inference Based on Ranks, New York: John Wiley & Sons, Inc, 1984

11 Nikitin Y. Asymptotic efficiency of non-parametric tests. Cambridge: Cambridge University Press, 1995

12 Vapnik V. Statistical Learning Theory. New York: Wiley, 1998

13 Joachims T. Making large-scale SVM learning practical. In: Schölkopf B ed. Advances in Kernel Methods-Support Vector Learning. California: MIT Press, 1999

(2003-07-14 收稿, 2004-03-16 收第 1 次修改稿, 2004-04-29 收第 2 次修改稿)