# 瘤诊断 的应用

健②

数学 1; ②Department of C er Science and Engineering, ew Yo USA. * 联系人, E-ma a@math.pku.edu.cn)

是当 中的一个重要研究 ，其中最主要的问 学中的 了秩和基因选取 并利用支持向量 行训练建 实验表明这种方法 模型可使得在结肠 别达到

持向量机 选取

基因表达 这些生物 识别提供 析出有 究的主

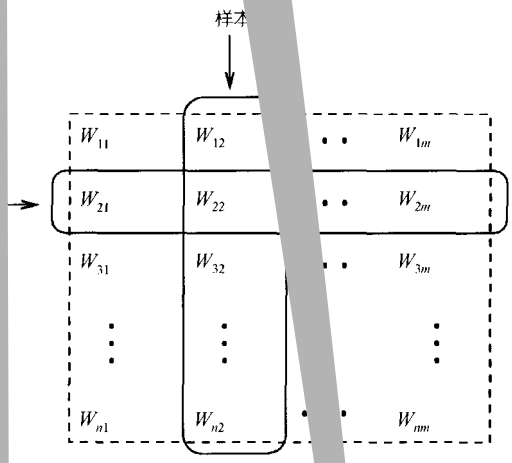


图 1 基因表达谱矩

因表 瘤分类 其常用的 . 特别地, 基 已成为其中 一个 [1]首先采用 邻域 过程中采用了一 量选取了 个 等[ 对结肠 的基因 表达谱与 瘤的 量 法进行 相关基 [3]将几种 用分 方法应 至 分类, 并对分类 果进 了 (SVM) 类最好. 这一结 n 等[6 研究结

很不理想. 主要表现在肿 分类方法的推 不足, 即根据训练样本集所得到的分类规则 本集上表现出较低的正确率, 即使采用推 好的 SVM 也是如此. 我们认为其主要症结 很好的剔除基因表达谱中的噪声. 实际中, 的出现可能仅仅与某些基因的表达水平的 . 若笼统地用全部基因表达水平来进行分 会因数据维数的巨大而难于进行, 而且众多 将便成噪声大大地干扰分类的结果. 为此 提出了一些相关基因选取的方法 应用最广泛

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

| | 结肠正常样本 | 结肠肿 | 据集的正 性检验结果 本 | 乳腺肿瘤样本 | 急性淋巴白血病 | 急性髓性白血病 |
|---|---|---|---|---|---|---|
| 数 | 2000 | 20 | | 5776 | 7129 | 7129 |
| 被否定的基因数 | 730 | 148 | | 574 | 4542 | 2558 |

算. 因此按照 A, B 两基因真实的 p 值按 率高 样本量之下，由秩和方法得到相有可能排在 A 基因的前面. 这就是 t 统 关基 肿瘤因素的辨识性.

条件下造成了基因的错位. 第二，当基因 谱 思想上是对每个基因的表达水平从正态分布时，计算出的 t 统计量也不再 从 t 做秩 肿瘤和正常样本的分布间是否存，因此无法通过查统计表或利用统计软件计算 在显 是，表明肿瘤因素对该基因的表达基因真实的 p 值. 得不到 p 值就不能利用 著性水 水平 因此 基因与肿瘤关系紧密，将阈值确定合适的相关基因数目，这将 中确定 其选 计量 直接利用基因表达水平适的相关基因数目带来不便. 因此 说，t 的值 法先 对表达水平排序得到计量方法只有在正态性条件满足 能表 “秩 位置)， 秩和统计量进行分得最好. 然而，正态性假设无 量方 析. 的基因 数据中含有大量的的应用范围. 噪声 噪声和 极大的影响 t 统

为了验证实际问题中基 态 计量 对利用 和统计量的值分布，我们在结肠数据 1) 据 造成 说“秩” 因此秩和上利用峰度和偏度[9] 样 方法 更适合于 因表达谱本服从正态分布，所 数据 一类错误的概率小 本类别数 分为出，约半数基因 Wilc Kruskal-W 秩和检验. 前者设未被否定的 用于 类别( 正常)的相关基因选因此显然不能 取问题，后者适用于多 的基因选取. 本文着从正态假设 两肿瘤类别的 问题，以下简述表达谱数 瘤分类问题的基因表达谱不服 W oxon 秩和检验用于 的基本步骤:

从正态分布的情况应该是普遍的. 在正态假设不满 (1) 建立假设

一般建立如下假设 和正常样本(或两

小到 序，排序号称为 同数值的用平均秩表 如，有 $k$ 个相 分别排在 $n+1$ 到 $n$ 则每一个表 都用 $n+\frac{k+1}{2}$ 表示.

统计量

一类 $n_1$ 表示，计算该类 $H_0$ 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 data
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 来自 http: me.wi.mit.edu/Mi ALL_AML.html
个正常样本，14 个肿瘤样本，5776 个基因. 来自 http://genc

文

结肠
白血

函数
Matl
light
M
hims
下，
选
然
的
VM
验
式
4

比较

| 数据集 | 基因选取 | 34 个基因 | 8 个基因 |
|---|---|---|---|
| 结肠数据 | 秩和方 | 96.2% | 88.8% |
| | t 统计量 | 93.6% | 88.8% |
| 数据集 | 基因选 | 844 个基因 | 398 个基因 |
| 白血病数据 | 秩和 | 100% | 98.6% |
| | t 统计 | 95.4% | 94.4% |

一优势还是十分显著的
结肠和白血病这两个数
正态分布，因此实验结
数据集上优于 t 统计量

## 3 总结

本文对于基于基
研究，而肿瘤诊断问
本文针对传统的相
足，提出了秩和基
正态条件的缺陷，
另一方面，秩和方
来确定出与肿瘤
数的确定更为科
来确定相关基因
机(SVM)在选取
诊断模型．实
SVM 的肿瘤
据和白血病数
率．实验也
法的．这些
诊断模型是

致谢 本工
助项目．

1999, 96: 6745~6750
P S, Grundy W N, Lin D, et al. Knowledge-based
microarray gene expression data by using support
hines. Proc Nat' l Acad Sci, 2000, 97(1): 262~267
Fridyand J, Speed T P. Comparison of discrimination
the classification of tumor using gene expression data.
merican Statistical Association, 2002, 97(457): 77~87
ristianini N, Duffy N, et al. Support vector machine
and validation of cancer tissue samples using
expression data. Bioinformatics, 2000, 16(10):

ston J, Barnhill S, et al. Gene selection for cancer
using support vector machine. Machine Learning,
389~422
ston J, Cai J, et al. Gene functional Analysis from
data. Proc Fifth Int. Conf. on Computational
gy. New York: ACM Press, 2001. 249~255
ysis of gene expression profiles: class discovery
In: Proc RECOMB, 2002. 127~136
thods of Statistical Analysis. (2nd edition). New
& Sons, 1956
. Statistical Inference Based on Ranks, New
Sons, Inc, 1984
otic efficiency of non-parametric tests.
University Press, 1995
arning Theory. New York: Wiley, 1998
rge-scale SVM learning practical. In:
es in Kernel Methods-Support Vector
ress, 1999

1 Gol
of
m

2 A

收稿, 2004-03-16 收第 1 次修改稿,
2004-04-29 收第 2 次修改稿)