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

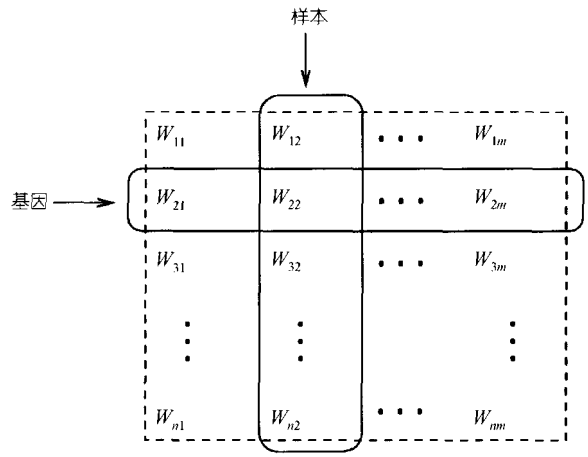

图 1 基因表达谱矩阵

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).

为了
验理
SVM
数据(
验样本
数据
可以
下文
方法
，并
节中
应用
t 统
结
铁
本

N．此时分类间隔等于 $2/\|w\|$，使分类间隔最

介于使 $\|w\|$ 最小．使 $f(w)=\frac{1}{2}\|w\|^2$ 最小且满足

的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题转化为其对偶问题，即在约束条件

$_i=0, a_i \geq 0, i=1,2,\cdots,N$ 下，求优化目标函

$)=\sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i,j=1}^{N} a_i a_j y_i y_j (x_i x_j)$ 的最大值．这是

不等式约束下的二次优化问题，存在惟一解．可

优分类面为 $f(x)=\mathrm{Sgn}\left(\sum_{i=1}^{N} a_i^* y_i (x_i x)+b^*\right)$，容

，此解中只有一小部分 a_i^* 不为零，其对应的样支持向量．b^* 是分类面的阈值，可以由支持向量

于线性不可分情况，可以通过非线性变换将化成高维特征空间中的线性问题，在特征空最优分类面．注意到，在上面的对偶问题中，优化目标函数还是分类函数都只涉及样本内比，只需用 $K(x,x')$ 代替原先的内积，即相当于题空间变换到新的特征空间．此时优化的目

变为：$Q(a)=\sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i,j=1}^{N} a_i a_j y_i y_j K(x_i,x_j)$，

应的分类函数也变为：$f(x)=$

$a_i * y_i K(x_i,x)+b^*\Big)$，其他约束条件不变．

对不同类型的数据，SVM 可以使用不同的核积形式)．目前比较常用的有：多项式核函数 $=[(xx_i)+1]^q$；径向基(Gauss)核函数 $K(x,x_i)=$

取出来的分类面．

图 2 支持向量机示意

使

y_i

诊断模型. 实验表明, 秩和基因选取方法及结合 SVM 的肿瘤诊断模型是有效的. 该诊断模型在 据和白血病数据上分别达到了 96.2%和 100% 率. 实验也证明了秩和方法是优于传统的 t 法的. 这些结果表明秩和方法及其结合 SVM 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: 助项目.

参 考 文 献

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

2 Alon U, Barkai N, Notterman D A, et al. Br expression revealed by clustering analysis colon tissues probed by oligonucleotide ar

8 Ding H Q. Analysis 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)