

基于非线性回归分析的差异基因选择方法

朱钦平 胡晓涵 祁云嵩

(江苏科技大学计算机科学与工程学院, 镇江 212003)

摘要 提出了一种新的用于微阵列基因差异表达多重假设检验的统计量计算方法, 该方法利用基因表达值到各类样本数据中心的距离作为统计量进行多重假设检验, 各统计量之间没有相关性, 并且有效地减弱了数据噪声带来的假阳性结果, 从而提高了多重假设检验的功效, 所选择出的基因集也具有更好的分类能力。

关键词 基因微阵列 基因选择 差异表达基因 多重假设检验

中图法分类号 TP391.4; **文献标志码** A

基因芯片(microarray)又称DNA芯片、DNA微阵列。它将设计好的探针, 有序高密度地在载体上排列成DNA微阵列, 待测样品与芯片杂交后通过扫描系统检测信号强度并对信号进行综合分析后, 获得样品中大量基因序列及表达信息^[1]。鉴于其快速、高通量、样品用量少等优点, 现正广泛地运用于生物学研究。但是该技术产生的数据由于其高噪音、高维度的特点, 在数据处理方面仍没有满意的解决方法^[2]。目前基因芯片在生物学研究中主要用于实现疾病的细分及预测病人的预后。前者多用无监督的算法实现, 主要为聚类分析, 在芯片实验中最为常用; 而后者多用有监督的算法, 包括判别分析、人工神经网络模型等方法^[3—5]。但在进行上述分析之前, 为了降低数据的维度和排除假阳性, 往往都需要筛选在不同条件下差异表达的基因。因此挑选出差异表达的基因往往是芯片数据分析的第一步。差异基因的选择问题可以看成是一类多重假设检验问题。所谓多重假设检验, 就是从手头样本出发, 针对某一个问题提出原假设(该假设是一系列假设, 并非单一假设), 按照统计规律对假设的正确与否进行推断。

多重假设检验用于对基因微阵列数据进行分

析处理进行基因选择, 即从中提取出据有样本分类价值的鉴别基因。目前, 基于多重假设检验的基因选择方法多采用t检验作为统计量^[6], 对于第j个基因的统计量取值计算如下: $t_j = \frac{\bar{x}_{1j} - \bar{x}_{2j}}{s_{1j} + s_{2j}}$, 其中, $\bar{x}_{1j}, \bar{x}_{2j}$ 表示两类样本中各类样本所有基因表达值平均值, 而 s_{1j}, s_{2j} 则分别表示丙类样本中各类样本所有基因表达值的均方差。全用这种统计量的缺陷是各基因的统计量之间是相互影响的, 这就违背了多重假设检验中各统计量是相互独立的前提, 同时, 利用这种统计量选择出来的基因, 无法排除噪声数据引起的统计量的显著性。本文提出了一个基于非线性回归分析方法用于微阵列数据的多重假设检验, 实验结果表明, 该方法选择出的基因集具有更好的分类特性。

1 基于非线性回归分析的多重假设检验

基于非线性回归分析方法用于微阵列数据的多重假设检验方法, 其基本原理是先对微阵列数据中各别的样本分别进行非线性回归分析, 得到各别样本的基因表达回归曲线, 然后计算待测样本中各基因表达值对于各样本回归曲线的距离, 用这些距离作为统计量进行基因表达差异性的假设检验。本文采用基于有理插值样条的非线性回归算法对微阵列数据进行回归分析。

2010年5月26日收到

第一作者简介: 朱钦平(1983—), 男, 硕士研究生, 研究方向: 模式识别与智能系统。E-mail: zqp_2000@126.com。

1.1 基于函数值的带参数有理样条

设在区间 $[a, b]$ 上给定插值点及函数值 (x_i, y_i) , $i = 1, 2, \dots, n$, 使得 $a = x_1 < x_2 < \dots < x_n = b$ 列成区间 $[a, b]$ 上的一个剖分。记 $h_i = x_{i+1} - x_i$, 对区间 $[a, b]$ 上的任一点 x , 令 $\lambda = (x - x_i)/h_i$ 。对任一 y_i , 构造在 $[x_i, x_{i+1}]$, $i = 1, 2, \dots, n-2$ 上的一元有理插值为

$$P_i^*(x) = p_i^*(x)/q_i^*(x), i = 1, 2, \dots, n-2 \quad (1)$$

式(1)中 $p_i^*(x) = (1 - \lambda)^3 \beta_i y_i + \lambda (1 - \lambda)^2 V_i^* + \lambda^2 (1 - \lambda) W_i^* + \lambda^3 y_{i+1}$, $q_i^*(x) = (1 - \lambda) \beta_i + \lambda$ 。并且, $V_i^* = (\beta_i + 1) y_i + \beta_i y_{i+1}$, $W_i^* = (\beta_i + 2) y_{i+1} - h_i \Delta_{i+1}^*$ 。这里, $\beta_i > 0$, $\Delta_i^* = (y_{i+1} - y_i)/h_i$ 。

这种插值被称之为基于函数值的一元有理插值, 它满足: $p_i^*(x_i) = y_i$, $p_i^*(x_{i+1}) = y_{i+1}$, $p_i^{*'}(x_i) = \Delta_i^*$, $p_i^{*'}(x_{i+1}) = \Delta_{i+1}^*$ 。

显然, 插值函数 $p_i^*(x)$ 在 $[x_i, x_{i+1}]$ 上对插值数据 (x_i, y_i) , $i = 1, 2, \dots, n$ 和参数 β_i 是唯一的。

1.2 样条逼近算法

一般一元非线性回归模型为

$$y = \varphi(x, \theta) + \varepsilon \quad (2)$$

式(2)中, $\theta = (\theta_1, \theta_2, \dots, \theta_r)$ 为未知参数向量, ε 为随机误差变量, $E(\varepsilon) = 0$, $\text{var}(\varepsilon) = \delta^2$ 。

采用上述基于函数值的一元有理插值样条对 $\varphi(x, \theta)$ 进行逼近, 即寻求有理插值样条:

$p_i^*(x) = \varphi_i(x, \theta)$, $i = 1, 2, \dots, n-1$, 确定参数 β 的具体值, 即可求得所需回归模型。具体算法如下:

(1) 从已知样本数据 $a = x_1 < x_2 < \dots < x_n = b$ 中挑选适当的数据组成新的样本 $(x_1^*, x_2^*, \dots, x_m^*)$, 一般令 $x_1^* = x_1$, $x_m^* = x_n$ 。计算 $h_i^* = x_{i+1}^* - x_i^*$, $i = 1, 2, \dots, m-1$ 。为了使所得函数为区间 $[a, b]$ 上的函数, 故需加入新点 $x_{m+1}^* = x_m^* + h_{m-1}$, 取 (x_{m-1}^*, y_{m-1}^*) , (x_m^*, y_m^*) 进行线性插值得插值函数 $y = kx + c$, 将 x_{m+1}^* 代入得 y_{m+1}^* , 由此得到新的样本点 (x_{m+1}^*, y_{m+1}^*) 。

(2) 对新的样本数据 $x_1^* < x_2^* < \dots < x_m^* < x_{m+1}^*$ 采用上述的一元有理插值样条方法进行插值。令 $\lambda = (x - x_i^*)/h_i^*$, 对每一段 $[x_i^*, x_{i+1}^*]$, $i = 1, 2, \dots, m-1$ 构造插值函数:

$$P_i^*(x) = p_i^*(x)/q_i^*(x), ? i = 1, 2, \dots, m-1.$$

(3) 取适当的 $\beta_i > 0$, $i = 1, 2, \dots, m-1$ 得到

$[x_i^*, x_{i+1}^*]$, $i = 1, 2, \dots, m-1$ 上的确定函数 $P^{**}(x) = P_i^*(x_i, \beta_i)$ 即为 $[a, b]$ 上的插值函数, 即观测数据 (x_i, y_i) , $i = 1, 2, \dots, n$ 的非线性回归方程。

1.3 构造新的统计量

先对两类样本的各基因表达值进行非线性回归分析, 得到两类样本的表达曲线 L_1 , 式 L_2 。则新的检验统计量 $D_j = d_{L_1, L_2}$, 其中, d_{L_1, L_2} 表示第 j 个基因的回归曲线 L_1 到 L_2 的位移。由于该假设检验统计量计算的是各基因表达值到不同的样本中心(非线性回归曲线)的距离, 因而各基因的统计量之间没有相关性, 并且有效地减弱了数据噪声带来的假阳性结果, 从而提高了多重假设检验的功效。

1.4 多重假设检验

假设 H_0 表示两组样本之间无显著差异, 观察 p 值表示样本总体大于随机变量观测值的几率。在多重假设检验中, 通常我们无法使用原始观察 p 值作为拒绝零假设的标准。为此, 需要采用合适的方法调整 p 值。

Benjamini and Hochberg 于 1995 年首先提出 FDR 概念及其控制程序⁽⁷⁾。FDR 概念如下: 同时对 m 个假设进行检验, 其中 m_0 个是正确的, R 表示检验结果为阳性的假设个数, V 表示拒绝假设 H_0 的个数, $FDR = \begin{cases} E(V/R), & R \neq 0 \\ 0, & R = 0 \end{cases}$, FDR 为拒绝 H_0 的结果中错误者所占比例之期望, 即“阳性结果错误率”。

FDR 控制: 同时检验 m 个假设: $H_{01}, H_{02}, \dots, H_{0m}$, 相应原始 p 值为: p_1, p_2, \dots, p_m , 令单个检验水平 $\alpha = 0.05$ 。BH 法步骤:(1) 把原始 p 值按从小到大的顺序排列 $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$, 相应零假设为 $H_{0(1)}, H_{0(2)}, \dots, H_{0(m)}$ 。(2) 从 $p_{(m)}$ 开始, 估计 $\hat{k} = \arg \max_{1 \leq k \leq m} \left\{ k : p_{(k)} \leq \frac{k}{m} \alpha \right\}$ 。(3) 若存在 \hat{k} , 则拒绝 k 之前的假设 $H_{0(1)}, H_{0(2)}, \dots, H_{0(k)}$, 反之接受所有的 $H_{0(i)}$ ($i \in R$ 且 $1 \leq i \leq m$)。相应的 p 值调整为 $\tilde{p}_{(i)} =$

$$\min_{k=i, i+1, \dots, m} \left\{ \min \left(\frac{m}{k} p_{(k)}, 1 \right) \right\}.$$

2 实验

为了检验基于非线性回归分析的多重假设检验,我们从 CNS 实验网站下载的公共数据集进行实验分析;其数据组成为 42 个肿瘤样本包括 10 个髓母细胞瘤,10 个横纹肌样脑膜瘤,10 个胶质瘤,8 个幕上原始神经外胚层肿瘤和 4 个正常人小脑,CNS 原始数据集用鲁棒多芯片平均(RMA)和 GC 鲁棒多芯片平均(GCRMA)进行了预处理。

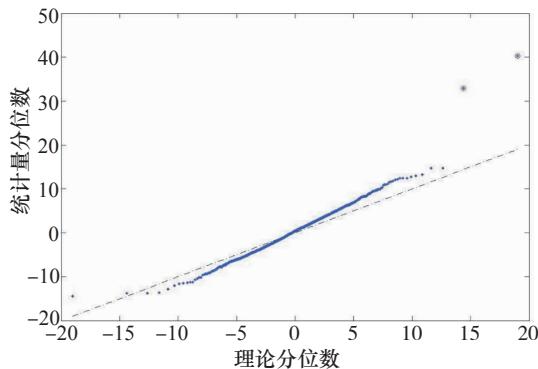


图 1 基于非线性回归距离的统计量正态分位图

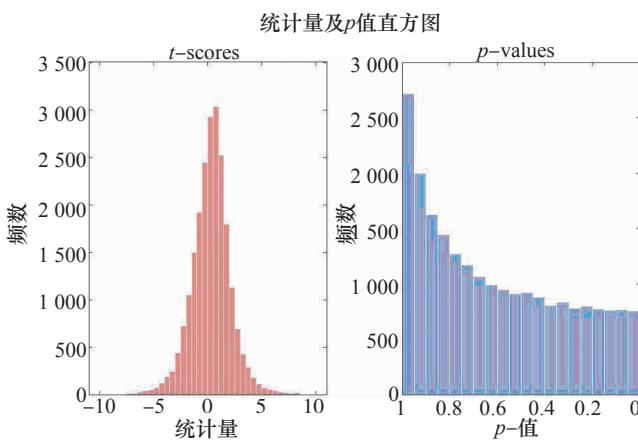


图 2 统计量及 p 值直方图

图 1 是基于非线性回归距离的统计量正态分位图,可以看出,该统计量是基本符合正态分布的,可以用于多重假设检验分析。图 2 是各基因统计量以及根据 Benjamini 和 Hochberg 于 1995 年首先提出

FDR 的概念及其控制算法^[7]计算出的各基因表达差异的 p 值分布图。

实验据中共有有效基因数 5 758,选择基因差异表达 p 值的阈值 0.05,则实验认定有鉴别能力的基因数为 923。而用于对用于对照的基于 t -统计量的方法选择基因数为 1 275。对两种方法选择出的基因集进行比较,本文研究的方法结果中,在 t -统计量的方法选择基因集中排列最前面的部基因落选,其原因是数据噪声带引起的假阳性结果。根据基因选择结果,实验采用了目前各研究论文中常用的分类效果公认较好的支撑向量机分类器对微阵列数据进行分类。基于 t -统计量的方法和本文的方法选择出的两个基因集下的分类准确率分别为 95% 和 97%。结果表明,本文方法选择出的基因集具有更好的分类能力。

4 结论

多重假设检验要求各统计量之间是相互独立的。对于基因微阵列数据,由于各基因间的相互作用,传统的基于 t -统计量的方法计算出的统计量之间能满足多重假设检验所要求的统计量之间的独立性要求。本文利用基因表达值到各类样本数据中心的距离作为统计量进行多重假设检验,各统计量之间没有相关性,并且有效地减弱了数据噪声带来的假阳性结果,从而提高了多重假设检验的功效,所选择出的基因集也具有更好的分类能力。

参 考 文 献

- Oshlack A, et al. Using DNA microarrays to study gene expression in closely related species. *Bioinformatics*, 2007; 23(10):1235—1242
- Zheng W, et al. Microarray-based method to analyze methylation status of E-cadherin gene in leukemia. *Clinica Chimica Acta*, 2008; 387(1-2): 97—104
- Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. *Genome Res*, 2001;11(11):1878—1887
- 李颖新, 阮晓刚. 基于支持向量机的肿瘤分类特征基因选取. *计算机研究与发展*, 2005;42(10):1796—1801
- 刘全金, 李颖新, 朱云华, 等. 基于 BP 神经网络的肿瘤特征基因选取. *计算机工程与应用*, 2005;41(34):184—186

- 6 Dudoit S, Shaffer J P, Boldrick J C. Multiple hypothesis testing in microarray experiments. *Stat Sci*, 2003;18:71—103
- 7 Benjamini Y, Liu W. A step down multiple testing procedure that controls the false discovery rate under independence. *J Statist Plann Interference*, 1999;82:163—170

Differential Expression Genes Selection Based on Nonlinear Regression Analysis

ZHU Qin-ping, HU Xiao-han, QI Yun-song

(Insistute of Computer Science & Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, P. R. China)

[Abstract] The new treatment method applied to multiply hypothesis statistic calculating methods of microarray differential genes expression is proposed. The method uses the distance between value of gene expression and various types of sample data center as statistic of multiply hypothesis testing. There is no correlation between all the statistics. The method also effectively reduced the data noise caused by false-positive. Therefore it improve the effectiveness of multiple hypothesis testing. The set of selected genes also has better classification.

[Key words] gene microarray gene selection differentially expressed genes multiple hypothesis testing

(上接第 6660 页)

Research and Simulation of A CCD Image Sensor with a Vertical Anti-blooming Strcttrue

WU Li-fan

(Xi'an University of Post & Telecommunications, Xi'an 710121, P. R. China)

[Abstract] MEDICI is a powerful device simulation program that can be used to simulate the behavior of MOS and semiconductor devices. A simulation grid is created by MEDICI. The substrate voltage, the 1 PW impurity concentration, N buried-channel impurity concentration and P impurity concentration of TG(transfer Gate) is analyzed. Finally, an optimum strctture is obtained.

[Key words] CCD blooming vertical anti-blooming device simulation