文章编号:1001-2486(2007)01-0111-05

基因表达缺失值的加权回归估计算法*

邱浪波[1,2],王广云[1,2],王正志[1]
(1. 国防科技大学 机电工程与自动化学院,湖南 长沙 410073; 2. 空军工程大学 电讯工程学院,陕西 西安 710077)

摘 要:在基因芯片实验中,数据缺失客观存在,并在一定程度上影响芯片数据后续分析结果的准确性。在不增加实验次数的情况下,缺失值估计是降低缺失数据对后续分析影响的有效方法。利用相似性信息的核加权函数来实现缺失值回归估计的局部化,提出了基于加权回归估计的基因表达缺失值估计算法。在两个不同类型的基因芯片数据上,将新方法与几种已知的方法进行了比较分析。实验结果表明,新的估计算法具有比传统缺失值估计算法更好的稳定性和估计准确度。

关键词:基因芯片表达;缺失值;加权回归
中图分类号:Q332 文献标识码:A

Missing Value Estimation for Microarray Expression Data Based on Weighted Regression

QIU Lang-bo[1,2], WANG Guang-yun[1], WANG Zheng-zhi[1]
(1. College of Mechatronics Engineering and Automation, National Univ. of Defense Technology, Changsha 410073, China;
2. Telecommunication Engineering Institute, Air Force Engineering Univ., Xi'an 710077, China)

Abstract: In microarray experiments, the missing value does exist and somewhat affects the stability and precision of the expression data analysis. Compared with increasing experiments, missing value estimating is preferred in reducing the influence of missing values on the post-processing. With the kernel weight based on similarly between target gene and sample genes, which localize missing value estimation, a new method based on weighted regression is presented. On the two real microarray expression datasets, the novel method was compared with several existing methods. Experimental results show that the novel method has better stability and precision than the existing methods that have been employed.

Key words: microarray expression; missing value; weighted regression

基因芯片是分子生物学、微电子学和信息学等学科交叉形成的一种新型生物技术,目前已经广泛应用于分子生物学、生物医学等研究领域,如 DNA 测序、基因调控网络和癌症检测等[1-3]。由于实验中存在很多变异来源,如样品培养不充分、测试中的图像污染等,基因表达数据矩阵通常含有缺失。在芯片数据的聚类算法中,涉及到基因表达谱的相似性度量,分析结果显示,缺失值有可能对聚类结果造成严重的影响[2]。同样,芯片数据分类中的一些常用方法,如主分量分析(PCA)和独立分量分析(ICA),支持向量机(SVM)等,也无法处理含有缺失值的数据集[2-11]。因此,缺失值的有效处理成为基因表达数据分析的重要预处理环节[3]。

近年来,出现了一些新的缺失值估计方法,如基于奇异值分解的插值法(Singular value decomposition impute,SVDimpute),最近邻插值法(K-nearest neighbor impute,KNNimpute),基于概率 PCA 的插值法(Bayesian Principal Component Analysis impute, BPCAimpute),基于高斯混合聚类的插值法(Gaussian Mixture Clustering and Impute, GMCimpute),最小二乘插值法(Local least squares impute, LLSimpute)[5-11]。基于全局数据的 BPCAimpute 和 GMCimpute 与基于局部数据的 LLSimpute 方法在稳定性和估计准确度上均优于 KNNimpute 和 SVDimpute 两种方法。最小二乘估计没有充分考虑用于

* 收稿日期:2006-09-03
基金项目:国家自然科学基金资助项目(60471003)
作者简介:邱浪波(1977—),男,博士生。

回归分析的样本基因与目标基因间的相似性信息。将相似性信息作为权重,用于基因表达缺失值估计,本文提出了基于核加权回归估计的缺失估计算法(Weighted Regression impute, WRimpute)。

1 缺失值估计方法描述

用矩阵 $\boldsymbol{G}\in R_{m\times n}$ 表示 m 个基因在 n 个实验中的表达数据矩阵:

$$\boldsymbol{G}=\begin{pmatrix}\boldsymbol{g}_1\\ \vdots\\ \boldsymbol{g}_m\end{pmatrix}=\begin{pmatrix}g_{11} & \cdots & g_{1n}\\ \vdots & \ddots & \vdots\\ g_{m1} & \cdots & g_{mn}\end{pmatrix} \tag{1}$$

其中,$m\gg n$,矩阵 $\boldsymbol{G}$ 的行向量 $\boldsymbol{g}_i$ 表示第 i 个基因在 n 个实验中的不同表达水平值。假设第 i 个基因在第 j 个实验中的表达值缺失,定义为 α,记为:

$$G(i,j)=g_{i,j}=\alpha \tag{2}$$

为了简化算法的描述,假设只有第一个基因在第一个实验中的表达水平是缺失的,即:

$$G(1,1)=g_{1,1}=\alpha \tag{3}$$

1.1 基因表达水平的相似性度量

与目标基因具有相似性表达的基因子集拥有较全部基因更好的数据结构,因此,在对目标基因的缺失值进行回归估计时,选用 k 个与目标基因具有最大相似性表达的候选基因作为回归样本。在相似性表达基因选择的过程中,通常采用 L_2 范数或者 Pearson 相关系数来度量相似性程度[5]。在计算候选基因 $\boldsymbol{g}_j$ 和含有缺失值的目标基因 $\boldsymbol{g}_1$ 的相似性程度时,忽略缺失值所在维的元素。下面给出基于 L_2 范数的相似性度量:

$$r(\boldsymbol{g}_j,\boldsymbol{g}_1)=\sqrt{\frac{\sum_{i=2}^{n}(g_{1i}-g_{ji})^2}{n-1}} \tag{4}$$

其中,$r(\boldsymbol{g}_j,\boldsymbol{g}_1)$ 表示基因 $\boldsymbol{g}_j$ 和 $\boldsymbol{g}_1$ 之间的相似性程度。在本文中采用 L_2 范数度量相似性。样本加权函数根据样本基因与目标基因的相似性度量赋予样本一个权值来描述。定义:

$$K_\lambda(g,g_1)=D\left(\frac{r(g,g_1)}{h_\lambda(g_1)}\right) \tag{5}$$

其中,$D(t)=\frac{3}{4}(1-t^2)$,$h_\lambda(\boldsymbol{g}_1)$ 是一个宽度函数(被 λ 标引)[12],它确定 $\boldsymbol{g}_1$ 的领域宽度,本文采用 $k-$最近邻域,即约束为与目标基因 $\boldsymbol{g}_1$ 具有最大相似性表达的前 k 个基因入选,k 取代了 λ,并且有 $h_k(g_1)=r(g_{s_k},g_1)$,其中 $\boldsymbol{g}_{s_k}$ 是第 k 个与目标基因 $\boldsymbol{g}_1$ 具有最大相似性表达的基因,从而实现了回归样本的局部化。

由上式得到样本权值:

$$w_{s_i1}=\frac{K_\lambda(g_{s_i},g_1)}{\sum_{s_j=1}^{k}K_\lambda(g_{s_j},g_1)} \tag{6}$$

计算目标基因 $\boldsymbol{g}_1$ 与所有候选基因的相似性度量,选择具有最大相似性表达的 k 个候选基因作为估计目标基因 $\boldsymbol{g}_1$ 缺失值的样本集,从而得到相似性表达基因集 G_s,以及相似性度量核权矩阵 W:

$$\boldsymbol{G}_s=\begin{pmatrix}\boldsymbol{g}_{s1}\\ \vdots\\ \boldsymbol{g}_{s_k}\end{pmatrix}=\begin{pmatrix}g_{s_11} & \cdots & g_{s_1n}\\ \vdots & \ddots & \vdots\\ g_{s_k1} & \cdots & g_{s_kn}\end{pmatrix}$$

和

$$\boldsymbol{W}=\begin{pmatrix} w_{s_1 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{s_k 1} \end{pmatrix} \tag{7}$$

1.2 加权最小二乘估计算法

核加权回归估计在目标点 g_1 解一个单独的加权最小二乘问题:

$$\min_{\beta(g_1)} \sum_{i=1}^{k} w_\lambda(g_{s_i}, g_1)[y_i - \beta(g_1) x_i] \tag{8}$$

令 $Q[\beta(g_1)] = [y - X\beta(g_1)]^{\mathrm{T}} \boldsymbol{W} [\boldsymbol{y} - \boldsymbol{X}\beta(g_1)]$,然后,对于 $\beta(g_1)$ 求导,并令结果等于零,可以得到

$$\hat{\boldsymbol{\beta}}(g_1) = (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W} y \tag{9}$$

选择 k 个最大相似性基因,得到相似性表达矩阵 $\boldsymbol{G}_s$。定义矩阵 $\boldsymbol{C} \in R_{k \times (n-1)}$,向量 $\boldsymbol{p} \in R_{1 \times (n-1)}$ 和 $\boldsymbol{d} \in R_{k \times 1}$,其中 $\boldsymbol{C}$ 由 $\boldsymbol{G}_s$ 中与目标基因 $\boldsymbol{g}_1$ 非缺失值对应的列向量构成,$\boldsymbol{d}$ 由 $\boldsymbol{G}_s$ 的第一列元素构成,$\boldsymbol{p}$ 由目标基因 $\boldsymbol{g}_1$ 的非缺失值元素构成。表述为:

$$\begin{pmatrix} \boldsymbol{g}_1 \\ \boldsymbol{G}_s \end{pmatrix} = \begin{pmatrix} \alpha & \boldsymbol{p}_{1 \times (n-1)} \\ \boldsymbol{d}_{k \times 1} & \boldsymbol{C}_{k \times (n-1)} \end{pmatrix} = \begin{pmatrix} \alpha & g_{12} & \cdots & g_{1n} \\ g_{s_1 1} & g_{s_1 2} & \cdots & g_{s_1 n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{s_k 1} & g_{s_k 2} & \cdots & g_{s_k n} \end{pmatrix} \tag{10}$$

在 WRimpute 缺失估计算法中,定义 $\boldsymbol{C}$ 为自变量的数据矩阵 $\boldsymbol{X}$,$\boldsymbol{d}$ 为因变量的数据向量 $\boldsymbol{y}$,则由式(9)可得到一个解向量 $\hat{\boldsymbol{\beta}}(g_1)$(回归系数向量)。然后利用 $\hat{\boldsymbol{\beta}}(g_1)$ 来估计目标基因 $\boldsymbol{g}_1$ 的缺失值。

$$\alpha = \boldsymbol{p}\hat{\boldsymbol{\beta}}(g_1) = (g_{12}, \cdots, g_{1n}) \hat{\boldsymbol{\beta}}(g_1) \tag{11}$$

为了能够充分利用全局数据,提高回归估计的稳定性,可采用一个迭代的步骤。当对所有缺失基因完成一次估计后,在下一次估计中,除了当前选定的目标基因以外,其余作为候选基因(排除与当前目标基因有相同缺失位置的基因),对目标基因的缺失位置进行重新估计,直到前后两次估计值的差异小于某个阈值。

1.3 参数 k 的估计

参数 k 对数据本身的结构特点有一定的依赖性,可以采用启发式方法得到一个近似最优的 k 值。通过与目标基因 g_1 的相似性度量和排序,得到一个有序的候选基因序列。假设目标基因 g_1 含有 s 个缺失值,删除这 s 个缺失值得到一个不含缺失的 $n-s$ 维全表达向量 q。同理,删除候选基因与目标基因 g_1 缺失值所在维对应的元素,得到一个 $n-s$ 维的候选基因全表达向量。假设 q 的某一维缺失,确定一个 k 值,得到候选基因序列中前 k 个维度为 $n-s$ 维的最大相似性候选基因集。此时,利用 WRimpute 方法对 q 的缺失维进行估计。由于 q 本身是不含缺失的,则可以对该 k 值的估计性能进行评估。对同一个 k 值,选择 q 不同的维缺失,重复实验,得到一个对 k 值更为可靠的性能评估。此时,可近似认为具有最佳性能的 k 值能最好地估计目标基因的缺失值。

2 实验验证

2.1 数据和评估方法

选择两组实验数据。数据一为研究 Saccharomyces cerevisiae 细胞周期调控信息的 Elutriation 芯片数据[1],包含 6718 个基因和 14 个实验 数据二是酵母环境反应实验芯片数据[13],包含 6361 个基因和 156 个实验,参照文献[9]的选择策略,与时序数据相对应,在对比实验中只采用含缺失较少且相关性较小的实验芯片数据,得到包含 15 个实验的芯片实验数据测试集。在两个测试数据中,第一个数据是时序芯片数据,第二个数据为非时序芯片数据。

为了估计缺失值估计算法的性能,将数据集中所有含有缺失值的基因删除,从而得到一个新的不含

缺失值的全表达矩阵。现在按照某个缺失百分比,随机从新全表达矩阵中移除元素,得到一个含有缺失值的表达矩阵。利用缺失估计方法,对此矩阵中的缺失值进行估计,最终得到一个估计矩阵。用估计值与原始值的标准化偏差来度量估计算法的性能。

$$NRMSE = \frac{\sqrt{\text{mean}(R_i - I_i)^2}}{\text{std}[I_i]} \tag{12}$$

其中,R_i 为估计值,I_i 为原始值,std$[I_i]$为原始值的偏差。

2.2 对比结果

对比实验中,分别在两个数据集上将 WRimpute 方法与当前比较好的三种方法 BPCAimpute、GMCimpute 和 LLSimpute 作了比较,其中后三种方法采用文献作者提供的源代码。两个数据集经过完全化预处理后,在缺失百分比分别为 1%和 5%的情况下,对 BPCAimpute、GMCimpute,以及 WRimpute 和 LLSimpute 方法在选取不同相似性表达基因个数 k 时的性能进行了测试。后两种方法在取同一 k 值时重复实验 50 次,取标准偏差 NRMSE 的均值作为每种方法在取某一 k 值时的性能评估值。BPCAimpute 采用 $n-1$ 个主成分,GMCimpute 采用最大 5 个聚类,BPCAimpute 和 GMCimpute 均重复实验 50 次。

从图 1 和图 2 均可以看出,随着 k 值的增加,WRimpute 和 LLSimpute 两种方法的性能明显优于 BPCAimpute 和 GMCimpute 性能,在一定程度上说明局部数据较全局数据有更好的数据分布结构。而且,在参数 k 一个较宽的区间范围内,WRimpute 和 LLSimpute 两种方法性能的变化比较平稳。可见,LLSimpute 和 WRimpute 两种方法对参数 k 都具有比较好的稳定性。在处理时序和非时序两种类型的数据中,WRimpute 在整体性能上均优于 LLSimpute。在图 1 和图 2 的对比中,还可以看到,由于时序芯片数据具有较好的相关性,在不同 k 值下,各种方法的估计精度均优于对非时序数据估计的精度。

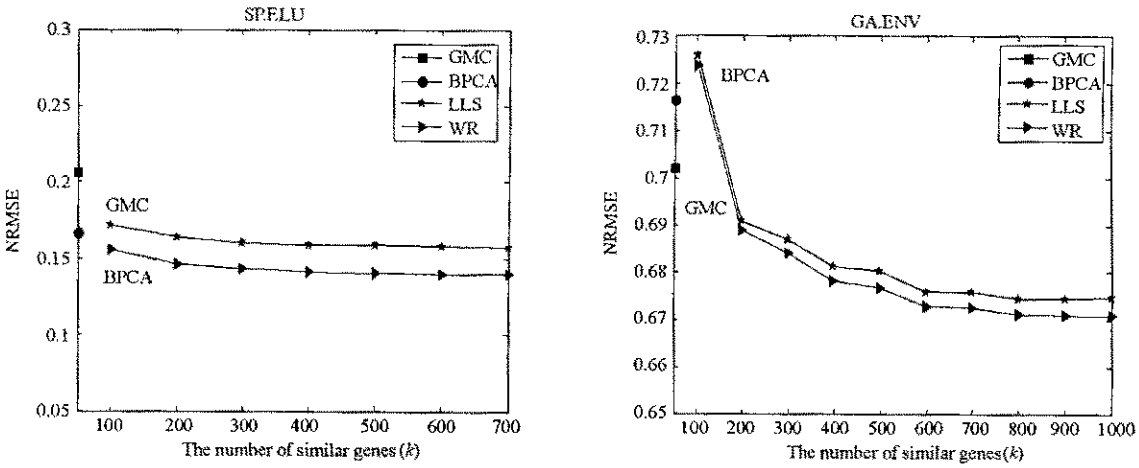

图 1 在缺失比为 1%的两个基因芯片数据上,几种方法在选取不同相似性基因数量时的 NRMSE 性能比较

Fig. 1 Comparison of the four methods and effect of the number of similar genes for estimating missing values on SP. ELU dataset and GA. ENV dataset those have 1% entries missing

同时,在不同缺失百分比的情况下,对几种方法性能变化的趋势进行了评估。数据分别取 1%~10%缺失百分比。BPCAimpute 采用 $n-1$ 个主成分,GMCimpute 采用最大 5 个聚类,WRimpute 与 LLSimpute 采用前述方法自动确定 k 值。在每个百分比,每种方法重复实验 50 次,取 NRMSE 的均值作为每种方法在取某一缺失百分比时的性能评估值。

由图 3 知,随着缺失比的增加,各种估计方法的性能均有下降。特别是在处理时序数据时,性能下降比较明显。基于局部回归估计的 WRimpute 和 LLSimpute 方法性能均优于 BPCAimpute 和 GMCimpute。在低缺失比情况下,WRimpute 和 LLSimpute 两者的性能比较接近,WRimpute 略优于

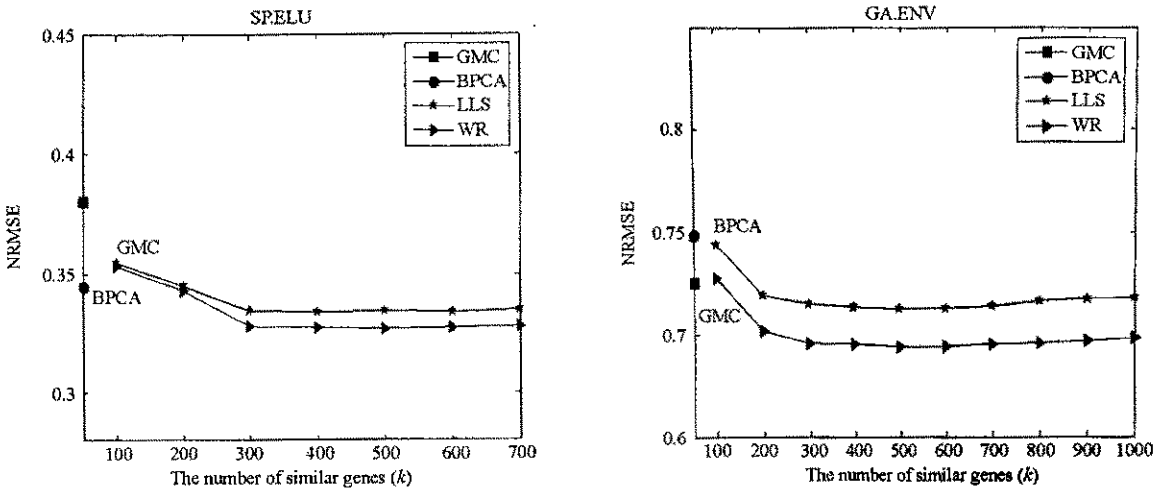

图 2 在缺失比为 5% 的两个基因芯片数据上,几种方法在选取不同相似性基因数量时的 NRMSE 性能比较

Fig. 2 Comparison of the four methods and effect of the number of similar genes for estimating missing values on SP. ELU dataset and GA. ENV dataset those have 5% entries missing

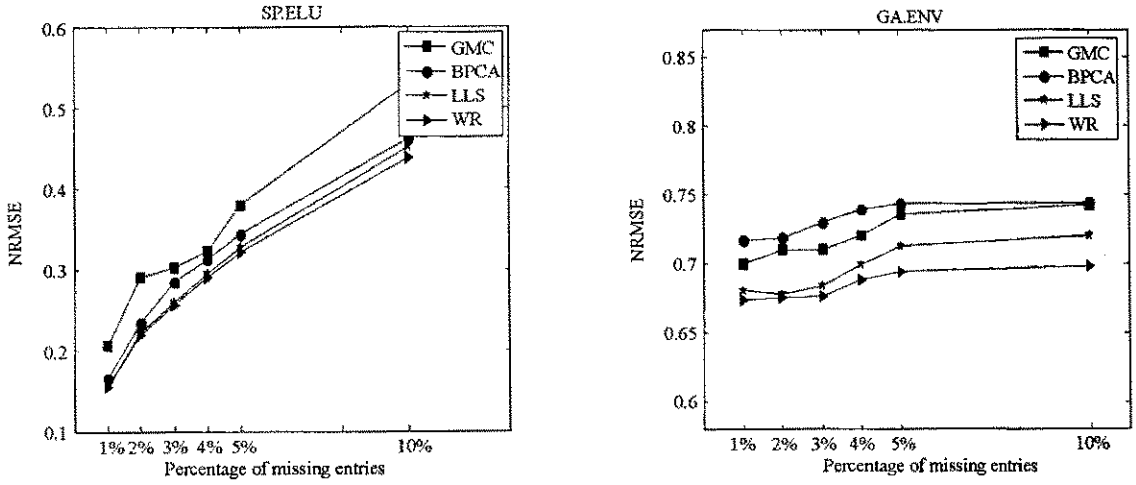

图 3 在取不同缺失百分比的情况下,四种缺失估计算法的 NRMSE 性能比较

Fig. 3 Comparison of the four methods against percentage of missing entries on SP. ELU dataset and GA. ENV dataset

LLSimpute。**在处理非时序数据的时候,** WRimpute **方法的优势更为明显。这表明局部数据具有较全局数据更好的数据结构,通过局部化处理能够有效提高缺失估计的精确度。**

3 结 论

有效的缺失值估计将减小缺失值对芯片数据后续分析的影响,提高后续分析的稳定性和准确性。利用基于相似性的核加权信息,提出了一种基于加权回归估计的缺失值估计算法 WRimpute。**实验结果表明,**WRimpute **方法具有更好的稳定性和估计准确度,为芯片数据缺失值的有效处理提供了一种新的方法,有助于在芯片数据的后续分析中得到更为准确的生物分析结果。**

(下转第 125 页)

$$K_2(t,s)=\begin{cases}1+(t-a)(s-a)+\frac{1}{6}(t-a)^2(3s-t-2a), & t\leqslant s\\ 1+(s-a)(t-a)+\frac{1}{6}(t-a)^2(3t-s-2a), & t>s\end{cases}$$

3 结束语

建立的再生核形式简洁、计算方便,并建立了直接的递推关系,为再生核的程序化计算提供了启示。另外,本文作者之一曾用空间正交分解和投影的方法研究过抽象算子样条和算子方程[5],而本文采用的内积使得分解式(2)是正交分解,这为将再生核与样条插值相结合的研究提供了基础。

参 考 文 献:

[1] Dalzell C J, Ramsay J O. Computing Reproducing Kernels with Arbitrary Boundary Constrains[J]. SIAM. J. Sci. Comput., 1993(14): 511-518.

[2] 崔明根,邓中兴. W_2^1 空间中的最佳插值逼近算子[J]. 计算数学,1986,8(2):209-216.

[3] 崔明根,吴勃英. 再生核空间数值分析[M]. 北京:科学出版社,2004.

[4] Schumaker L. Spline Functions: Basic Theory[M]. New York: John Wiley & Sons, 1981.

[5] 张新建. 希氏空间中算子样条插值及算子方程的近似解[J]. 数学学报,2002,45(2):227-234.

(上接第 115 页)

参 考 文 献:

[1] Spellman P T, Gavin S, Michael Q Z. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray HybridizationhD[J]. Molecular Biology of the Cell, 1998, 9: 3273-3297.

[2] Alexandre G B, Serge H, Alain M. Influence of Microarrays Experiments Missing Values on the Stability of Gene Groups by Hierarchical Clustering [J]. BMC Bioinformatics, 2004, 5: 114-123.

[3] 李瑶. 基因芯片与功能基因组[M]. 北京:化学工业出版社, 2004.

[4] Liebermeister W. Linear Modes of Gene Expression Determined by Independent Component Analysis[J]. Bioinformatics, 2002, 18(1), 51-60.

[5] Olga T, Michael C, Sherlock G, Pat B, Trevor H, Robert T, David B, Russ B A. Missing Value Estimation Methods for DNA Microarrays[J]. Bioinformatics, 2001, 17: 520-525.

[6] Sandrine D, Jane F, Terence P S. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data [J]. Journal of the American Statistical Association, 2002, 97: 77-87.

[7] Zhou X B, Wang X D, Edward R D. Missing-value Estimation Using Linear and Non-linear Regression with Bayesian Gene Selection[J]. Bioinformatics, 2003, 19(17): 2302-2307.

[8] Oba S. A Bayesian Missing Value Estimation Method for Gene Expression Profile Data[J]. Bioinformatics, 2003, 19: 2088-2096.

[9] Ouyang M, Welsh W J, Georgopoulos P. Gaussian Mixture Clustering and Imputation of Microarray Data[J]. Bioinformatics, 2004, 20 (6): 917-923.

[10] Bo T H, Dysvik B, Jonassen I. LSimpute: Accurate Estimation of Missing Values in Microarray Data with Least Squares Methods[J]. Nucleic Acids Res, 2004, 32(3), e34.

[11] Hyunsoo Kim, Gene H G, Haesun P. Missing Value Estimation for DNA Microarray Gene Expression Data: Local Least Squares Imputation [J]. Bioinformatics, 2005, 2(21): 187-198.

[12] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning, Data Mining, Inference, and Prediction[M]. 范明,柴玉梅, 等译. 北京:电子工业出版社, 2004.

[13] Gasch A P, Spellman P T, Kao C M, et al. Genomic Expression Responses to DNA-damaging Agents and the Regulatory Role of the YeastATRhomolog Mec1p[J]. Mol. Biol. Cell, 2001, 12: 2987-3003.