文章编号:1004-9037(2013)02-0136-05

时序基因表达缺失值的加权双向回归估计算法

李建更 郭庆雷 贺益恒

(北京工业大学人工智能与机器人研究所,北京,100124)

摘要:由于受实验条件等客观因素制约,实验所得到的基因表达谱数据存在数据缺失的现象,不利于数据的后续使用。如何在不丢失数据信息、不影响数据整体使用的情况下,对实验数据进行估计、填充已成为目前生物信息学研究的热点。本文通过利用核加权函数提取与缺失值所在的行列具有最大相似性的行列信息,提出了基于双向核加权回归估计的算法。在回归过程中同时考虑基因表达的空间相关性和时间相关性信息,使回归算法使用的信息更加充分。通过与其他缺失值估计算法相比较,加权双向回归算法的估计结果较好。

关键词:时序基因表达;空间相关性;时间相关性;加权双向回归;缺失值估计

中图分类号:Q786;O212;TP391 **文献标志码**:A

Double Weighted Regression Estimation for Missing Values in Time Series Gene Expression Data

Li Jiangeng, *Guo Qinglei*, *He Yiheng*

(Institute of Artificial Intelligence and Robots, Beijing University of Technology, Beijing, 100124, China)

Abstract: Due to the limited experimental condition, there are missing values in gene expression data which make the following use difficult. Estimating missing values without data destroy and information lost has become an important work of bio-information. By weighted kernel function, it can find out rows and columns having largest similar coefficient with the rows and columns containing missing values. An estimation method based on double weighted regression is introduced by using weighted kernel function. It makes the information data more abundant by considering gene space correlation and time correlation in regression. Comparing with other methods, the weighted double regression method can obtain better estimation result.

Key words: sequential gene expression; space correlation; time correlation; double weighted regression; missing value estimation

引 言

对基因表达谱数据进行深入的分析与挖掘,获取其中潜在的生物过程信息,是构成基因表达谱数据分析的重要内容。但是受环境及实验条件等客观因素的影响,实际获取的基因表达谱数据存在着不同程度的缺失[1]。如果简单地将包含缺失数值的基因或者样本删除,将可能造成部分有用信息丢失,数据挖掘结果发生错误或者不完整。缺失值的填充成为对基因表达谱数据进行预处理的重要一环[2]。

对于缺失值的估计填充,已有许多文献 进行了相关研究,并提出修正方法。如基于奇异值分解(Singular value decomposition, SVD)、K 最近邻方法(K-nearest neighbor algorithm, KNN)、行平均值(Row average, RA)、最小二乘插值法(Local least square impute, LLSI)等[3-5]。这些方法能对常用基因表达谱数据缺失值进行有效估计,并取得理想的效果[6]。

对于时序 DNA 表达谱中的缺失数据,大多采用的是回归的方法进行估计。文献[7]通过对基因

基金项目:北京市教育委员会科技计划(JC002011200903)资助项目;水体污染控制与治理重大专项——南水北调中线总干渠水质安全保障关键技术与工程示范(2009ZX07212-003)资助项目。

收稿日期:2011-11-23;**修订日期**:2012-05-30

表达中的独立分量的分析建立了线性模型，有效地降低了缺失值填充过程中的计算复杂度。文献[8]比较了一元线性回归方法、多元线性回归方法及迭代回归方法对时序 DNA 表达谱数据的估计结果，发现利用与缺失值有较高相关性的基因，一元线性回归方法可以取得较好的估计效果；文献[9]利用自回归模型对 DNA 微阵列时序数据缺失值进行估计；文献[10]利用线性动态系统模型对缺失值进行了估计。

文献[11]提出了基因表达缺失值的加权回归估计算法，但是其中仅考虑基因的位置相关性，并且其算法核心为通过核方法选择与目标基因相关的基因并进行多元线性回归得到缺失值的估计值。本文在其基础上考虑基因表达值的时间相关性，在进行空间线性回归的同时加入了时间线性回归，提出了基于核加权的双向回归填充(Weighted double regression impute，WDRimpute)算法。

1　缺失值估计方法

1.1　时序基因表达谱数据

设时序基因表达谱数据 $\boldsymbol{G}=\{\boldsymbol{g}_1,\boldsymbol{g}_2,\boldsymbol{g}_3,\cdots,\boldsymbol{g}_N\}$，$\boldsymbol{g}_i$ 表示一个样本在时序点 i 时所有基因表达值，$|\boldsymbol{G}|=N$ 表示基因表达过程中存在 N 的检测时间点，$|\boldsymbol{g}_i|=M$ 表示一个样本的基因的数量 M。时序基因表达矩阵可以表示如下

$$\boldsymbol{G}=\begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,N} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ g_{M,1} & g_{M,2} & \cdots & g_{M,N} \end{bmatrix} \tag{1}$$

通常情况下，$M\gg N$，即基因的数量远大于采样的时间点数，呈现出基因表达谱数据的典型特征：小样本、高维度。假设基因表达谱数据的缺失值 $\boldsymbol{G}(i,j)$ 为空，即第 i 条基因在 j 时刻的表达值缺失。

1.2　最大表达相似性度量的基因选择

在对缺失值进行估计的过程中，回归样本选用与缺失值所在的基因具有最大相似度量的 S 个基因向量及与缺失值所在的时序列具有最大相似度量的 T 个时序列。通常采用 L_2 范数相关系数度量相似程度[4]。假设第一个基因位点在第一个时间点的表达值为缺失值，即 $\boldsymbol{G}(1,1)=\alpha$ 缺失。

行向量相似性度量

$$l(\boldsymbol{g}_j,\boldsymbol{g}_1)=\sqrt{\frac{\sum_{i=1}^{N}(g_{1,i}-g_{j,i})^2}{N-1}} \tag{2}$$

列向量相似性度量

$$l(\boldsymbol{g}_i,\boldsymbol{g}_1)=\sqrt{\frac{\sum_{j=1}^{M}(g_{j,1}-g_{j,i})^2}{M-1}} \tag{3}$$

根据上述相似性度量函数构建样本加权函数。定义如下

$$K_\lambda=D\Big(\frac{l(\boldsymbol{g},\boldsymbol{g}_1)}{h_\lambda(\boldsymbol{g}_1)}\Big) \tag{4}$$

式中 $D(t)$ 为 Epanechniv 二次核函数

$$D(t)=\begin{cases}\dfrac{3}{4}(1-t^2) & |t|\leqslant 1\\ 0\end{cases} \tag{5}$$

$h_\lambda(\boldsymbol{g}_1)$ 为宽度函数(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 $\boldsymbol{g}_1$ 具有最大相似性度量的前 K 个基因。对于行向量，$K=S$；对于列向量，$K=T$。并且 $h_\lambda(\boldsymbol{g}_1)=h_k(\boldsymbol{g}_1)=l(\boldsymbol{g}_{t_s},\boldsymbol{g}_1)$，$\boldsymbol{g}_{t_s}$ 是第 k 个与 $\boldsymbol{g}_1$ 具有最大相似性的基因，实现回归样本的局部化。

样本权值表示如下

$$\boldsymbol{w}_{s_i}=\frac{K_\lambda(\boldsymbol{g}_{s_i},\boldsymbol{g}_1)}{\sum_{j=1}^{S}K_\lambda(\boldsymbol{g}_{s_j},\boldsymbol{g}_1)} \tag{6}$$

$$\boldsymbol{w}_{t_i}=\frac{K_\lambda(\boldsymbol{g}_{t_i},\boldsymbol{g}_1)}{\sum_{j=1}^{T}K_\lambda(\boldsymbol{g}_{t_j},\boldsymbol{g}_1)} \tag{7}$$

根据前面得到的 S 个行向量与 T 个列向量，分别组成相似性基因表达数据集 $\boldsymbol{G}_S$，$\boldsymbol{G}_T$ 及相应的相似性度量核矩阵 $\boldsymbol{W}_S$，$\boldsymbol{W}_T$

$$\boldsymbol{G}_S=\begin{bmatrix}\boldsymbol{G}_{S1}\\ \vdots\\ \boldsymbol{G}_{SS}\end{bmatrix}=\begin{bmatrix} g_{S1,1} & \cdots & g_{S1,N}\\ \vdots & \ddots & \vdots\\ g_{SS,1} & \cdots & g_{SS,N}\end{bmatrix} \tag{8}$$

$$\boldsymbol{G}_T=(\boldsymbol{G}_{T1}\quad\cdots\quad\boldsymbol{G}_{TT})=\begin{bmatrix} g_{T1,1} & \cdots & g_{T1,T}\\ \vdots & \ddots & \vdots\\ g_{TM,1} & \cdots & g_{TM,T}\end{bmatrix} \tag{9}$$

和

$$\boldsymbol{W}_S=\begin{bmatrix} w_{S1} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & w_{SS}\end{bmatrix} \tag{10}$$

$$\boldsymbol{W}_T=\begin{bmatrix} w_{T1} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & w_{TT}\end{bmatrix} \tag{11}$$

1.3　加权最小二乘回归估计算法

核加权回归估计在目标点 $\boldsymbol{g}_1$ 解一加权最小二

乘问题[12]

$$\min_{\beta(\boldsymbol{g}_1)}\sum_{i=1}^{k}\boldsymbol{w}(\boldsymbol{g}_i,\boldsymbol{g}_1)[\boldsymbol{y}_i-\beta(\boldsymbol{g}_1)\boldsymbol{x}_i] \tag{12}$$

式中 k 为 S 或 T。结果为

$$\hat{\beta}(\boldsymbol{g}_1)=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{13}$$

对于其中 $\boldsymbol{X}$,$\boldsymbol{W}$ 与 $\boldsymbol{y}$,均使用样本矩阵与向量中的相关元素即可,如下所示

$$\begin{bmatrix}\boldsymbol{g}_{1s}\\ \boldsymbol{G}_S\end{bmatrix}=\begin{bmatrix}\alpha & \boldsymbol{P}_{1\times(N-1)}\\ \boldsymbol{y}_{(S-1)\times1} & \boldsymbol{X}_{(S-1)\times(N-1)}\end{bmatrix} \tag{14}$$

$$\begin{bmatrix}\boldsymbol{g}_{1t} & \boldsymbol{G}_T\end{bmatrix}=\begin{bmatrix}\alpha & \boldsymbol{y}_{1\times(T-1)}\\ \boldsymbol{P}_{(M-1)\times1} & \boldsymbol{X}_{(M-1)\times(T-1)}\end{bmatrix} \tag{15}$$

式中:$\boldsymbol{g}_{1s}$,$\boldsymbol{g}_{1t}$ 分别代表缺失值所在的行向量与列向量。代入式(13),可得缺失值所在的行向量与列向量分别对应的回归解向量:$\hat{\beta}_s(\boldsymbol{g}_{1s})$,$\hat{\beta}_t(\boldsymbol{g}_{1t})$。

然后,利用回归解向量估计目标基因 $\boldsymbol{g}_1$ 缺失值

$$\alpha_s=\boldsymbol{P}_{1\times(N-1)}\hat{\beta}_s(\boldsymbol{g}_{1s}) \tag{16}$$

$$\alpha_t=\hat{\beta}_t(\boldsymbol{g}_{1t})\boldsymbol{P}_{(M-1)\times1} \tag{17}$$

缺失值最终的估计填充结果为

$$\alpha=\lambda\cdot\alpha_s+(1-\lambda)\cdot\alpha_t,0\leqslant\lambda\leqslant1 \tag{18}$$

1.4 参数 S,T 的选择

参数 S,T 的选择对矩阵的结构有很大的依赖性,尤其是 T 的选择。由于客观条件的限制,时序点比基因点少很多。一般选择除去含有缺失值之外的所有的列向量。S 则使用 $1<S<2T$ 之间的数据。利用和加权回归方法对目标基因的非缺失值部分进行回归估计。重复实验,得到估计效果最好时的 S 值。

1.5 参数 λ 的选择

对于 λ 的计算,首先选择具有最大相似性系数的行及列向量,将行与列向量的交叉点的值作为假想缺失值。使用其他具有最大相似性的行列向量进行双向回归估计,得到相应的行与列回归值。之后利用式(18)计算 λ 的值,在式(18)中 α 为原始值,α_s,α_t 分别为对应的行、列回归值。根据最大相似性的定义,可以近似地认为在缺失值的处理过程中 λ 的值与此时的 λ 值相近[3]。

2 双向加权回归算法步骤

对第 1 节的算法进行整理,双向回归加权算法的步骤如下:

(1) 计算缺失值所在的行列和其他不含缺失值的行列之间的相似性度量。

(2) 选择前 S,T 个具有最大相似性度量行列。

(3) 根据式(4)计算步骤(2)选择的行列向量的加权函数,并基于加权函数计算行列向量的相似基因矩阵 $\boldsymbol{G}_S$,$\boldsymbol{G}_T$ 及对应的相似度量核矩阵 $\boldsymbol{W}_S$,$\boldsymbol{W}_T$。

(4) 解式(12)得到具有最大相似性度量的行列向量与缺失值所在的行列之间对应的回归解向量$\hat{\beta}_s(\boldsymbol{g}_{1s})$,$\hat{\beta}_t(\boldsymbol{g}_{1t})$。

(5) 选择与缺失值所在的行列具有最大相似性度量的行列交叉点的数据作为与缺失值具有最大相似性度量的值,重复步骤(1~4),根据式(16~18),得到与之对应的回归解向量,并计算参数 λ。

(6) 将步骤(5)得到的 λ 应用到缺失值对应的回归解向量,得到最终的缺失值对应的估计值。

3 实验验证

3.1 实验数据与评估方法

本文采用的数据为 NCBI 数据库中的 Lung development 数据集,包含 285 157 个基因位点 11 个时间点的表达数据。由于实验条件的限制,选择标准差小于 90 的 48 个基因位点表达值。在数据集中随机移除 $x(1\leqslant x\leqslant5)$ 个表达值,缺失百分比小于 5%。

实验结果的评价采用估计值与实际值的标准化偏差度量(Normalized RMS error, NRMSE)。

$$\mathrm{NRMSE}=\frac{\sqrt{\mathrm{mean}(R_i-I_i)^2}}{\mathrm{std}[I_i]} \tag{19}$$

式中:R_i 为估计值;I_i 为实际值;std[I_i]为实际值的标准差[2]。

3.2 实验结果及对比

在实验中,随机在数据集上选择 1~5 个数值作为缺失值,利用加权双向回归估计算法进行填充估计,同时利用 KNN 算法及缺失值对应的行列平均值,具有最大相似性表达基因行列平均值对缺失值进行填充,对 4 种填充算法的结果标准化偏差进行比较。

在不同缺失值个数情况下,采取不同的具有最大相似性的行向量个数 S 和列向量个数 T,分别进行回归填充,并记录对应的标准化偏差 NRMSE,得到 125 组数据。

图 1 中 SimRow$i$ 中的 i 指的是具有最大相似性度量的行向量的数量,SimCol$i$ 则表示 i 个具有最大相似性度量的列向量。

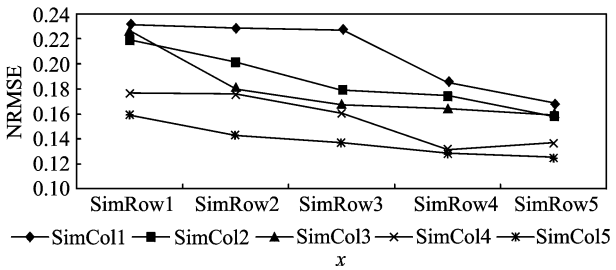

图 1 不同数目的具有最大相似度量的行列对最终标准化偏差度量的影响

图 1 显示了不同数目的具有最大相似性的行及列向量对标准化偏差度量的影响。可以看出，当具有最大相似度量的行向量固定后，随着列向量的数量的增加，标准化度量偏差逐渐降低。同样，当具有最大相似度量的列向量数量固定时，随着行向量的增加，标准化度量偏差逐渐降低。

随着相似的行列向量的增加，标准化偏差从最高的 0.2 以上逐渐降到了最低的 0.13 左右，即随着相似行列向量的数量的增加，数据填充值与实际值之间的标准化偏差逐渐降低。图 2 对图 1 的结果进行了总结。并且当具有最大相似度量的行列向量的数量相等时，回归得到的填充值大致相等。随着数量的变化，得到的两条曲线大致相同，而双向加权回归估计得到的值总是偏向于数据比较小的方向。这样回归得到的结果中包含了时间方向的加权回归结果和空间方向加权回归的结果。并且与单纯的行或者列加权回归结果相比，标准化方差变化幅度较小，即双向加权回归得到的结果与实际值之间的误差较为稳定。

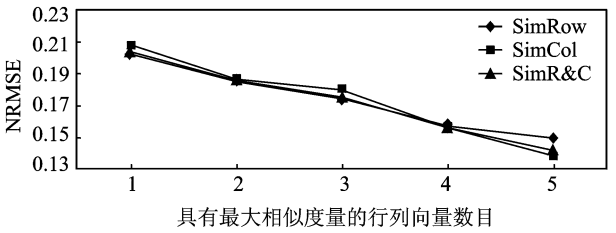

图 2　不同数目的具有最大相似度量的行或列对最终标准化偏差度量的影响

图 3 是不同数目时的缺失值估计结果。其中 DeleNum$i$ 表示 i 个需要进行填补的数值，4 种缺失值估计算法分别为 WDRimpute、K 最近邻填充算法（K-nearest neighbor impute，KNNimpute）、行列均值填充算法（Mean value of rows and columns impute，R&CMimpute）及行加权回归填充算法（Weight row regression impute，WRRimpute）。可以看出，当缺失值数目越多时，回归的标准化误差成下降趋势。而且，双向回归的结果一直优于其他 3 种方法。当缺失值数目达到 5 时，双向回归的标准化误差降到了 0.1 附近，行加权回归算法估计值的标准化误差达到 0.2 左右。可以看出，双向回归填充值的精度远高于其他填充方法得到的结果。

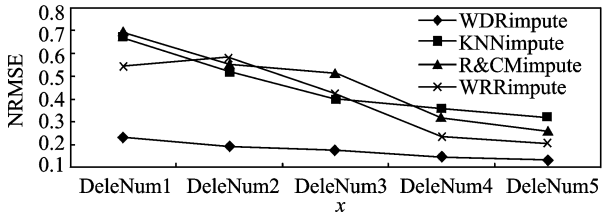

图 3　4 种不同的缺失值填充算法对不同数目缺失值的填充结果

4　结束语

本文基于文献[11]的加权回归算法，考虑到时序基因表达过程中的时间和空间因素的影响，提出了时序基因表达缺失值的双向加权回归估计算法，并将该方法应用于实际基因表达数据中。实验结果表明，利用双向回归估计算法得到的缺失值估计值与原数据的偏差明显小于其他填充值算法的偏差值，具有更好的估计精度，为以后生物基因表达谱缺失数据的处理提供了一种新的方法。需要注意的是，在总体数据量比较少或者需要进行估计的缺失值数量比较少的情况下，缺失值整体的估计误差可能比较大，使用的时候需要谨慎。

参考文献：

[1] 黄德双. 基因表达谱数据挖掘方法研究[M]. 北京：科学出版社，2009.

[2] 乔珠峰，田凤占，黄厚宽，等. 缺失数据处理方法的比较研究[J]. 计算机研究与发展，2006，43(S1)：171-175.
Qiao Zhufeng, Tian Fengzhan, Huang Houkuan, et al. A comparison study of missing value datasets processing methods [J]. Journal of Computer Research and Development, 2006, 43(S1): 171-175.

[3] 李序颖. 基于空间自回归模型的缺失值插补方法[J]. 数理统计与管理，2005，24(3)：45-50.
Li Xuying. Imputation method for regional missing data using spatial autoregression models[J]. Application of Statistics and Management, 2005, 24(3): 45-50.

[4] Olga T, Michael C, Sherlock G, et al. Missing value estimation methods for DNA microarrays [J]. Bioin-

formatics,2001,17:520-525.

[5] Sridevi S, Rajaram Dr S, Parthiban C, et al. Imputation for the analysis of missing values and prediction of time series data[C]// IEEE-International Conference on Recent Trends in Information Technology. [S. l.]:IEEE, 2011:1158-1163.

[6] 刘星毅. 一种杂合的缺失值填充方法[J]. 科技信息, 2007(27):418-420.
Liu Xingyi. A heterozygous imputation method of missing data[J]. Science & Technology Information, 2007(27):418-420.

[7] Liebermeister W. Linear modes of gene expression determined by independent component analysis [J]. Bioinformatics,2002,18(1):51-60.

[8] Jiang Yi, Lan Tuo, Wu Lihua. A comparison study of missing value processing methods in time series data mining[C]// CiSE 2009. New Jersey,USA:IEEE Computer Society,2009.

[9] Miew Keen Choong, Charbit M, Hong Yan. Autoregressive-model-based missing value estimation for DNA microarray time series data[C]// IEEE Transaction on Information Technology in Biomedicine. New Jersey, USA: Institute of Electrical and ELectronics Engineers Inc., 2009: 131-137.

[10] Connie Phong, Raul Singh. Missing value estimation for time series microarray data using linear dynamical systems modeling[C]// IEEE International Conference on Advanced Information Networking and Application. New Jersey, USA: Institute of Electrical and Electronics Engineers Inc., 2008:814-819.

[11] 邱浪波,王广云,王正志. 基因表达缺失值的加权回归估计算法[J]. 国防科技大学学报,2007,29(1):111-115,125.
Qiu Langbo, Wang Guangyun, Wang Zhengzhi. Missing value estimation for microarray expression data based on weighted regression[J]. Journal of National University of Defense Technology, 2007, 29(1):111-115,125.

[12] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning, data mining, inference, and prediction [M]. Berlin, Germany: Springer-Verlag, 2003.

作者简介:李建更(1965-),男,博士,副教授,研究方向:模式识别、机器学习、生物信息学;郭庆雷(1985-),男,硕士研究生,研究方向:模式识别、生物信息学,E-mail:guofeng2068@163.com;贺益恒(1987-),男,硕士研究生,研究方向:模式识别、生物信息学。