·综　　述·

基因表达谱缺失数据填补估计方法的研究进展与探讨*

伍亚舟 综述，易　东 审校
（第三军医大学卫生统计学教研室，重庆 400038）

关键词：基因表达谱；缺失数据；多重填补；支持向量回归
doi：10.3969/j.issn.1671-8348.2014.14.050　　**文献标识码**：A　　**文章编号**：1671-8348(2014)14-1806-03

基因芯片能为基因组学研究提供海量的基因表达谱数据，这些数据反映了基因在不同组织细胞的不同生长发育阶段或不同生理状态下表达水平的变化[1-2]。但是，由于基因表达谱的海量性、复杂性、噪声性和高维性特点，特别是缺失数据值的大量存在[3-5]，给后续的数据分析带来了较大困难，也产生了一些问题：观察到的数据与缺失数据间的差异所产生的偏倚，严重影响后续分析结果的客观性和正确性，从而导致后续分析质量的可靠性和稳健性降低，使得整个分析效率降低；另外，由于缺失数据的存在，经常得出难以解释的结果。因此，如何根据基因表达谱数据信息的特性进行有效的缺失值估计与填补是生物数据分析中重难点，并对后续基因表达谱的不同分析目的（如差异表达基因筛选、基因功能聚类、肿瘤组织分类）将产生重要的生物学影响[6-9]。本文针对基因表达谱缺失数据的特性，就当前国内外基因表达谱缺失数据的处理方法进行简要概述，在分析其各自优缺点基础上，提出并探讨一种新的填补估计方法。

1　基因表达谱缺失数据的产生原因与特性分析

微阵列数据通常以大规模矩阵的形式存在，该矩阵表示某个基因在不同试验条件（列）下的基因表达水平（行），但在实际情况中，实验获得的数据阵列通常是不完整的，即含有缺失值。缺失数据产生原因有多种，包括：不充分的实验方案，图像损坏，芯片上的灰尘或划痕等；另外，用来制造芯片的机械也可能系统地产生缺失数据。

事实上，基因表达谱缺失数据的缺失产生机制（完全随机缺失、随机缺失和非随机缺失）、缺失模式（单调缺失和任意缺失）、数据集序列类型（时间序列型、非时间序列型和混合序列型）、缺失率大小等特性，以及后续不同分析目的及其填补分析方法的要求，都会对缺失值填补与估计的准度产生较大影响。

2　常用的填补估计方法及其特点

缺失数据的估计与填补是在不增加实验次数情况下降低缺失数据对后续分析影响的有效方法。近年国内外学者在缺失数据的估计方面进行了有益的探索：(1)将存在缺失数据的行（基因）或实验条件（列）简单地从矩阵中剔除，以得到一个完备的数据集，称之为列表式删除；(2)直接在缺失数据集上进行数据挖掘，或利用一个特定的缺省值来填补；(3)利用统计学方法进行填补估计[3-5,9-16]：行均值，K 近邻法（KNN），奇异值分解（SVD），贝叶斯 PCA（BPCA），高斯混合聚类（GMC），最小二乘（LLS），支持向量回归，加权回归估计，极大似然估计（MLE），多重填补（MI）等。

2.1　常用填补估计方法

2.1.1　行均值法　实验表明，具有相似功能的基因在相同的微阵列杂交实验中会产生相似的表达模式。因此，依实验序列，同类中的基因表达模式极为相似，某个基因在某些条件下的缺失值，用缺失数据所在行的其他条件下的数据的平均值进行填补估计，即为行均值法。该方法简单易行，但并没有考虑数据间的关联性，其估计的准确度大大受影响。

2.1.2　K 邻近法　K 近邻法基本思路：首先计算每一个含有缺失值的基因和所有其他基因的欧式距离；在计算过程中，如果在同一个实验条件下两个基因有一个具有缺失值，则这个实验条件就不参与欧式距离的计算；再根据所计算得到的具有缺失值的基因和其他基因的欧式距离，选取和它最近的 K 个基因，Brettingham-Moore 等[1]分析发现 K 选取 10～20 比较合理。通过如下公式计算得到待补的缺失值：

$$W_i = D_i^{-1}\left(\sum_{k=1}^{k} D_k^{-1}\right)^{-1} \qquad (1)$$

$$G = \sum_{k=1}^{k} W_k G_I \qquad (2)$$

D_i 表示基因 G 与第 i 个近邻基因的欧式距离，W_i 表示为第 i 个近邻基因的权重，G_i 表示第 i 个近邻基因的表达值。G 通过 KNN 法计算得到的填补的缺失数据值。

2.1.3　马氏距离法　马氏距离方法是在 KNN 法基础上，通过基因之间的马氏距离来选择最近邻居基因，并将已得到的估计值应用到后续的估计过程中，然后采用信息论中熵值的概念计算最近邻居的加权系数，其相应位置的加权平均值即为缺失数据的估计值。该方法不仅考虑了观测变量之间的相关性，而且也考虑到了各个观测指标取值的差异程度，能更好地描述基因之间的相似程度。

2.1.4　随机回归填补法　随机回归填补是由单元的缺失项对观测项的回归，用预测值代替缺失值。通常由观测变量及缺失变量都有观测的单元进行回归计算。填补中还可以给填补值增加一个随机成分。它是用回归填补值加上一个随机项，预测出一个缺失值的替代值，该随机项反映所预测的值的不确定性影响。该方法能够较好的利用数据提供的信息，解决因预测变量高度相关引起的共线性问题。

2.1.5　极大似然估计法　极大似然估计法是在总体分布类型已知情况下的一种参数估计方法。在模型假定正确的情况下，若缺失机制为随机缺失，通过已观测数据的边际分布可以对未知参数进行极大似然估计，得到未知参数的准确估计值。该方

* **基金项目**：国家自然科学基金资助项目（81273178）。　**作者简介**：伍亚舟（1977-），副教授，博士，主要从事卫生统计学与生物信息学研究。

法需要有足够大的样本保证得到似然估计值是无偏的；另外，似然函数是基于完整数据某个假定的参数模型。实际应用中，如果模型假定错误，基于似然法的估计可能稳定也可能不稳定。

2.1.6 多重填补法 多重填补法由 Stekhoven 等[17]首先提出，该方法已被越来越多地应用于生物医学、统计学和机器学习等领域[18-20]。与单一填补(SI)的不同之处在于，MI 方法对每一个缺失值用某一可能值的集合进行填补，重复 p 次，故叫多重填补，从而产生若干个完整数据集；然后，用针对完整数据集的统计方法对每一个填补数据集分别进行统计分析，把得到的结果进行综合，进而产生最终的统计推断。

MI 方法的推断原理及主要步骤：首先，采用适当的填补方法模型，为每个缺失数据值产生一套可能的填补估计值，这些值反映了缺失值的不确性；每一个值都被用来填补数据集中的缺失值，产生若干个完整数据集(p 次)；其次，用针对完整数据集的统计方法对每一个填补数据集进行统计分析，得到每个缺失数据的均值和方差；最后，对来自于各个填补数据集的结果(缺失数据的均值和方差)以某种方法进行综合，从而产生最终的统计推断结果。

在 MI 出现以前，列表式删除和 SI 法是处理缺失值的主要方法，但是它们没有考虑到缺失数据的不确定性以及缺失数据与观察到的数据间可能存在的系统性差异，所以难以提供关于总体参数的准确估计。MI 弥补了单一填补和列表式删除等方法的缺陷，该方法能够反映出由于数据缺失造成的统计推断结果的不确定性，优化了多重填补方法的置信区间和相对效率。

2.2 常用填补估计方法的不足 基因表达谱缺失数据估计方法进展较快，但还存在许多难点和问题：(1)目前，很多估计方法多是 SI，即用一个可行的估计值对缺失数据进行一次填补，其优点是简单、速度快，适合于缺失率较低的表达谱数据，缺点是导致标准误降低和 P 值减小，使得犯Ⅰ类错误的概率升高，容易引起系统偏倚，且不能反映缺失数据值的不确定性，因此，用 SI 法计算出的治疗效应置信区间会失去它本来的真实性；(2)一些填补方法的应用条件相对较苛刻(如 KNN 法受变量类型限制，通常只适用于连续型变量[2]；(3)零或行均值法等没有考虑到数据本身的属性和数据间的相互联系；(4)直接删除会消除大量有效基因信息或使某个类消失，严重影响到后续分析结果的客观性和正确性。

3 基于支持向量回归的非参多重填补新融合方法

MI 方法虽然有无法替代的优点，但也有其缺陷。一方面，MI 在应用时，假设缺失机制是随机缺失，这种假设可以很方便地避开一些复杂的概率模型；另一方面，目前的具体多重填补模型参数方法都是要求数据集的分布已知，且对数据集的要求更为严格，如完整性、正态性和方差齐性等，实际上，由于在真实基因表达谱数据集中往往具有复杂数据结构，很难也几乎不可能精确地预测出缺失数据和可观测数据的关系，而且对将要处理的数据集没有任何先验知识。参数填补模型方法对此就束手无策或效果并不理想，而非参数模型方法在对数据分布未知的情况下却能取得很好的效果，比如基于核函数选择的支持向量机方法并结合回归分析的技术。因此，作者提出一种基于核函数的支持向量回归的非参多重填补(SVR-NPMI)的新融合方法，对基因表达谱缺失数据进行填补。

SVR-NPMI 方法将支持向量机和回归分析融合于多重填补的过程中，对缺失数据集进行多次填补(p 次)，最后利用参数和非参数统计方法进行综合估计，以达到填补缺失数据的目的。该方法中有两个问题需要注意：(1)填补次数 p 的确定要根据 γ(γ 为对总体参数缺失的部分信息的估计)来确定；(2)具体多重填补模型方法的确定，对于单调缺失模式，如针对连续型变量的预测均数匹配法和趋势得分法，针对离散型变量的判别分析和 Logistic 回归；对于复杂的缺失模式，可以采用马尔科夫链蒙特卡罗方法方法。

简要介绍基于 SVR 的非参多重填补融合方法的基本原理：

设某个非线性可分的基因表达谱数据集：

$$G=\{(x_1,z_1),(x_2,z_2),\cdots\cdots,(x_m,z_m)\} \quad (3)$$

这里 xi(i=1,2,…,m，m 为基因个数)为第 i 个基因的表达输入值，zi 为第 i 个基因的对应的目标输出值。

引入核函数 K，

$$K(x_i,x_j)=\phi(x_i)^T\phi(x_j) \quad (4)$$

常用的核函数有线性核、多项式核、高斯核、径向基核和 sigmoid 核等，核函数可以根据数据集的分布进行选择，从而达到最佳的效果。

于是 ε-支持向量回归可以表示为如下最优化问题：

$$\begin{cases} \underset{w,b,\xi,\xi^*}{min} & \frac{1}{2}W^TW+C\sum_{i=1}^{m}\xi_i+C\sum_{i=1}^{m}\xi_i^* \\ s.t. & W^T.\phi(x_i)+b-z_i\leqslant\varepsilon+\xi_i \\ & z_i-W^T.\phi(x_i)-b\leqslant\varepsilon+\xi_i^* \\ & \xi_i,\xi_i^*\geqslant 0,i=1,2,\cdots\cdots,m \end{cases} \quad (5)$$

其中 C 表示正则化参数，用来对模型复杂度和训练误差进行折中。引入拉格朗日乘子 α 和 α^*，将支持向量回归的原始问题转化为它的对偶形式：

$$\begin{cases} \underset{w,b,\xi,\xi^*}{min} & \frac{1}{2}(\alpha-\alpha^*)^TK(\alpha-\alpha^*)+\varepsilon\sum_{i=1}^{m}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{m}Z_i(\alpha_i-\alpha_i^*) \\ s.t. & \sum_{i=1}^{m}(\alpha_i-\alpha_i^*)=0 \\ & 0\leqslant\alpha_i,\alpha_i^*\leqslant C \end{cases} \quad (6)$$

于是支持向量回归的决策函数可以表示为如下表达式：

$$f(x)=\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)K(x_i.x)+b \quad (7)$$

若令 $\beta_i=\alpha_i^*-\alpha_i$，并令 $\beta=(\beta_1,\beta_2,\cdots\cdots\beta_m)^T$，则上式即为：

$$f(x)=\sum_{i=1}^{m}\beta_iK(x_i.x)+b \quad (8)$$

在上述每个原始数据集 G 中，在不包含缺失数据的基因中，以随机化原则抽取不同的基因数 $k(k\leqslant m)$构建训练数据集 G_{train}(p 个)进行训练，从而对包含缺失数据的基因构成的测试数据集 G_{test} 进行测试，得到最后的填补数据的估计值 $f(x)$，从而实现了缺失数据的预测。

4 结论与展望

本文针对基因表达谱缺失数据的特性，就当前国内外基因表达谱缺失数据的处理方法进行简要综述，在分析其各自优缺点基础上，提出并探讨一种新的填补估计方法——SVR-NPMI。该方法将多重填补、基于核函数选择的 SVM 和回归分析有机地融合在一起，具有明显优点：(1)弥补了 SI 的缺陷，

该法能够反映出由于数据缺失造成的统计推断结果的不确定性，优化了 MI 的置信区间和相对效率；(2)将 SI 与 MI 综合运用、参数与非参数统计方法相结合，使得新的融合方法受到数据分布的限制性更小、应用性更为广泛，可以解决表达谱数据本身的缺陷等问题；(3)该方法以与目标基因具有较高相似性的完全基因子集为训练集使用 SVR 算法(该算法具有非线性和鲁棒性，适于求解这种非线性的估计值问题)建立回归模型对缺失值进行估计，提高估计的准确性和稳定性，为基因表达谱缺失数据值的有效填补提供一种全新的思路方法。

在后续研究中，将利用基因表达谱公共数据集和自实验室数据集，证实基于 SVR-NPM 法对基因表达谱缺失数据进行估计的可靠性和有效性，建立一种基于不同序列数据集、不同分析目的、不同缺失率等情况下的缺失填补策略，并进一步阐明缺失填补方法对基因表达谱后续不同分析目的的生物学影响。

参考文献：

[1] Brettingham-Moore KH, Duong CP, Heriot AG, et al. Using gene expression profiling to predict response and prognosis in gastrointestinal cancers-the promise and the perils[J]. Ann Surg Oncol, 2011, 18(5): 1484-1491.

[2] Lee WP, Tzou WS. Computational methods for discovering gene networks from expression data[J]. Brief Bioinform, 2009, 10(4): 408-423.

[3] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays[J]. Bioinformatics, 2001, 17(6): 520-525.

[4] Dorri F, Azmi P, Dorri F. Missing value imputation in DNA microarrays based on conjugate gradient method[J]. Comput Biol Med, 2012, 42(2): 222-227.

[5] Little R, Rubin D. Statistical analysis with missing data [M]. New York: John Wiley and Sons Inc, 1987.

[6] Oh S, Kang DD, Brock GN, et al. Biological impact of missing-value imputation on downstream analyses of gene expression profiles[J]. Bioinformatics, 2011, 27(1): 78-86.

[7] Celton M, Malpertuy A, Lelandais G, et al. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments [J]. BMC Genomics, 2010, 11(1): 15-30.

[8] Sun Y, Braga-Neto U, Dougherty ER. Impact of missing value imputation on classification for DNA microarray gene expression data--a model-based study[J]. EURASIP J Bioinform Syst Biol, 2009, 2009: 504069.

[9] Oba S, Sato MA, Takemasa I, et al. A bayesian missing value estimation method for gene expression profile data [J]. Bioinformatics, 2003, 19(16): 2088-2096.

[10] Ouyang M, Welsh WJ, Georgopoulos P. Gaussian mixture clustering and imputation of microarray data[J]. Bioinformatics, 2004, 20(6): 917-923.

[11] Kim H, Golub GH, Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation[J]. Bioinformatics, 2005, 21(2): 187-198.

[12] Wang X, Li A, Jiang Z, et al. Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme[J]. BMC Bioinformatics, 2006, 7(1): 32-35.

[13] Berthoumieux S, Brilli M, de Jong H, et al. Identification of metabolic network models from incomplete high-throughput datasets[J]. Bioinformatics, 2011, 27(13): i186-i195.

[14] Tuikkala J, Elo L, Nevalainen OS, et al. Improving missing value estimation in microarray data with gene ontology[J]. Bioinformatics, 2006, 22(5): 566-572.

[15] 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.

[16] 杨涛，骆嘉伟，王艳，等. 基于马氏距离的缺失值填充算法[J]. 计算机应用，2005，25(12)：2868-2871.

[17] Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data[J]. Bioinformatics, 2012, 28(1): 112-118.

[18] Ryan R, Vernon S, Lawrence G, et al. Use of Name recognition software, census data and multiple imputation to predict missing data on ethnicity: application to Cancer registry records[J]. BMC Med Inform Decis Mak, 2012, 12(1): 1-8.

[19] Habbous S, Chu KP, Qiu X, et al. The changing incidence of human papillomavirus-associated oropharyngeal Cancer using multiple imputation from 2000 to 2010 at a Comprehensive Cancer Centre[J]. Cancer Epidemiol, 2013, 37(6): 820-829.

[20] Fong DY, Rai SN, Lam KS. Estimating the effect of multiple imputation on incomplete longitudinal data with application to a randomized clinical study[J]. J Biopharm Stat, 2013, 23(5): 1004-1022.

(收稿日期：2013-10-08 修回日期：2014-02-18)

医学论文中容易混淆的词语

箭头后为正确用字：

1%饿酸→1%锇酸
5－羟色氨→5－羟色胺
阿酶素→阿霉素
阿斯匹林→阿司匹林
记数法→计数法
甲氨喋呤→甲氨蝶呤
节段性肠炎→局限性肠炎
禁忌症→禁忌证
水份→水分
丝裂酶素→丝裂霉素
松弛→松弛
苔盼蓝→锥虫蓝