文章编号: 1001- 2486(2009) 01- 0124- 05

一种基于 KNN-SVR 的基因表达缺失值的估计方法*

王广云, 倪青山, 邱浪波, 王正志
(国防科技大学 机电工程与自动化学院, 湖南 长沙 410073)

摘 要: 为了消除不相似基因对基因表达谱中缺失值估计的影响, 提出了一种基于 KNN-SVR 的缺失值估计方法。该方法先通过最近邻法选出与目标基因表达最相似的一组完全基因, 再用这些基因通过支持向量回归对缺失值进行估计。还提出了用标准化偏差的方差来度量算法的稳定性和估计值的可信度。该方法通过对基因的过滤提高了缺失值估计的有效性。实验结果表明, KNN-SVR 法具有较高的估计精度和稳定性。

关键词: 基因芯片; 缺失值估计; 最近邻法; 支持向量回归; 相似性
中图分类号: Q332 文献标识码: A

Missing Value Estimation for Microarray Expression Data Based on KNN-SVR

WANG Guang-yun, NI Qing-shan, QIU Lang-bo, WANG Zheng-zhi
(College of Mechatronics Engineering and Automation, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: In order to exclude the effect of dissimilar genes, a new missing value estimation method based on KNN-SVR is proposed. This method selects a group of complete genes most similar to target genes by K-nearest neighbor (KNN) and uses them to estimate missing values by Support Vector Regression (SVR). This paper also suggests using the variance of Normalized Root Mean Squared Error (NRMSE) to measure the stability of estimation methods and the reliability of estimated values. This method improves the validity of missing value estimation by filtering genes. The experiment results show that KNN-SVR method has better accuracy and stability.

Key words: microarray; missing value estimation; K-nearest neighbor; support vector regression; similarity

随着基因芯片技术在生物学领域的广泛应用, 基因表达数据分析成为一个非常重要的环节[1]。然而, 由于表达数据存在一定的缺失, 使得许多数据分析算法不能对含有缺失的数据进行处理[2-5], 所以, 研究者提出了一些缺失值估计算法, 如最近邻法(K-nearest neighbor, KNN)[2]、贝叶斯主成分分析法(Bayesian Principal Component Analysis, BPCA)[3]和局部最小二乘法(Local Least Squares, LLS)[4]。但这些方法没有充分利用基因表达谱数据, 估计精度不高[1,5]。支持向量回归(Support Vector Regression, SVR)作为一种新型的统计学习方法, 在缺失值估计中得到了广泛的关注[5-6]。但是 SVR 使用了表达谱中所有基因, 建模时搜索空间大, 计算时间长, 而且在回归过程中, 与缺失基因相似性弱的基因会降低估计精度。

针对上述问题, 本文提出了一种将 KNN 法和 SVR 法相结合的缺失值估计方法。该方法首先通过 KNN 筛选出与目标基因相似的 K 个完全基因, 构成搜索空间, 再用 SVR 对目标基因的缺失值进行估计。同时, 还提出了用标准化偏差(Normalized Root Mean Squared Error, NRMSE)的方差来度量算法的稳定性, 在一定程度上反映了估计值的可信度。在对三组真实的基因表达谱数据的测试实验中, 证明了基于 KNN-SVR 的缺失值估计方法可以有效地提高缺失值的估计精度, 而且具有较高的稳定性。

1 KNN 法和 SVR 法

设 $A=[a_{ij}]_{m\times n}$ 表示 $m\times n$ 基因表达矩阵, 行为基因 $g_i(i=1,\cdots,m)$, 列为样本 $s_j(j=1,\cdots,n)$, a_{ij}

* 收稿日期: 2008- 09- 01
基金项目: 国家自然科学基金资助项目(60471003)
作者简介: 王广云(1980—), 女, 博士生。

为基因 g_i 在样本 s_j 中的表达值。设数据 a_{td} 缺失, 则 g_t 为目标基因, 在样本 s_d 中没有缺失的基因为完全基因。

1.1 KNN 法

KNN 法根据目标基因 g_t 与所有完全基因间的相似度, 选出与 g_t 最相似的 K 个完全基因, 然后按照下述公式计算得到待补的缺失值:

$$\omega_k = R_{tk} / \left(\sum_{k=1}^{K} R_{tk} \right), \hat{a}_{td} = \sum_{k=1}^{K} \omega_{tk} a_{kd} \tag{1}$$

其中, R_{tk} 表示 g_t 与第 k 个完全基因 g_k 的相似度, ω_{tk} 表示 g_k 的权重, $\hat{a}_{td}$ 为 a_{td} 的估计值。

KNN 算法简单, 没有充分考虑到数据间的相互关系, 精确度不高[1-5]。

1.2 SVR 法

设每个基因对应样本空间中的每个样本点, 则训练样本集为 $\{(x_i, y_i), i = 1, 2, \cdots, l, l \leqslant m\}$, 其中 $y_i = a_{id}$, $x_i = (a_{i1}, \cdots, a_{id-1}, a_{id+1}, \cdots, a_{in})$ 表示完全基因 g_i 中除去 s_d 后其他样本表达值所构成的向量。

SVR 的基本思想是: 以 x_i 为输入值, y_i 为对应的目标值, 将样本点用函数 $\Phi(x)$ 映射到高维特征空间进行线性回归, 从而获得在原空间中的估计结果。设估计函数为

$$f(x) = w \cdot \Phi(x) + b \tag{2}$$

则最优化问题为

$$\begin{aligned} &\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} (\xi + \xi^*) \\ &\text{s.t. } y_i - w \cdot \Phi(x_i) - b \leqslant \varepsilon + \xi \\ &\quad w \cdot \Phi(x_i) + b - y_i \leqslant \varepsilon + \xi_i^* \\ &\quad \xi_i \geqslant 0 \\ &\quad \xi_i^* \geqslant 0 \end{aligned} \tag{3}$$

其中, $C > 0$ 为惩罚系数, ξ 和 ξ^* 为松弛变量, w 的维数为特征空间的维数, $\varepsilon > 0$ 为与估计精度直接相关的设计参数。对偶最优化问题为

$$\begin{aligned} &\max_{\alpha, \alpha^*} \left\{ -\frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) - \varepsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i (\alpha_i - \alpha_i^*) \right\} \\ &\text{s.t. } \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0 \\ &\quad 0 \leqslant \alpha_i \leqslant C \\ &\quad 0 \leqslant \alpha_i^* \leqslant C \end{aligned} \tag{4}$$

最后得到回归估计函数为

$$f(x) = \sum_{x_r \in SV} (\alpha_i - \alpha_i^*) K(x_r, x) + b \tag{5}$$

其中, $K(x_r, x) = \Phi(x_r) \cdot \Phi(x)$ 为核函数。因此, 缺失数据 a_{td} 的估计值为 $\hat{a}_{td} = f(x_t)$。

SVR 可以得到唯一的、全局最优解, 而且 SVR 的解具有稀疏性, 通过改变 ε 可以控制解的稀疏性和估计的精度, 这就使得系统具有很强的泛化能力和实用性[6]。

2 KNN-SVR 估计方法

从 KNN 法的原理中可以看出, 用来对缺失值进行估计的 K 个基因对估计结果起着决定性的作用。同样, SVR 中的训练集样本对于估计结果也有很大的影响。以所描述的三个数据集(Data1/Data2/Data3)含 1% 的缺失数据为例, 就基因的相似性对 KNN 法和 SVR 法的影响分别进行了研究, 结果如表 1 所示。

表 1 列出了 KNN 法和 SVR 法在不同相似度基因对缺失值估计时的 50 次随机实验的 NRMSE 的平均值。NRMSE 的值越小表示估计精度越高。可以看出, KNN 法在 $K = 15$ 且 15 个基因分别为最相似的

15 个基因(15 similar)时的估计精度最高;在最相似的 8 个和最不相似的 7 个基因(8 & 7)上的精度次之;在最不相似的 15 个基因(15 dissimilar)上的精度最低。SVR 法也遵循同样的规律。实验结果表明,与缺失基因相似度比较低的基因会降低估计精度。

表 1 在 KNN 法和 SVR 法中不同相似度基因对缺失值估计的 NRMSE

Tab. 1 NRMSE of missing value estimation with different similar genes under KNN and SVR

Dataset	KNN			SVR		
	15 similar	15 dissimilar	8 & 7	20 similar	20 dissimilar	10 & 10
Data1	0.3291	1.0957	0.5001	0.2713	2.6873	0.5655
Data2	0.3473	1.1283	0.5875	0.3302	3.2466	0.5889
Data3	0.8227	1.1364	0.8987	0.7925	1.2109	0.9400

由于基因表达谱数据不符合均一的数据分布,可类似看作是多个分段函数的集合。因此,与目标基因具有较高相似度的基因子集拥有比全部基因更好的数据结构,通过局部化处理可以去除较低相似度基因的干扰,得到更高精度的估计值。针对上述问题,本文提出了一种将 KNN 与 SVR 相结合的方法。该方法采用 KNN 的思想,根据基因间的相似度,选择与目标基因最相关的 K 个完全基因作为训练集,然后再使用 SVR 方法对缺失值进行估计。具体步骤如下:

Step 1: 以 Pearson 系数[1] 为度量方法,计算一个目标基因与所有完全基因的相似度;

Step 2: 筛选出与该目标基因最相似的 K 个完全基因构成训练集 X_b;

Step 3: 以径向基函数[6] 为核函数,用 SVR 对该目标基因中的缺失值进行估计;

Step 4: 对其他目标基因进行上述的估计,直到估计出全部缺失值为止。

3 实验结果及分析

进行三组实验:第一组是人类结肠癌数据[7](Data1),包含 6500 个基因和 62 个样本;第二组是人类白血病数据[8](Data2),包含 7129 个基因和 72 个样本;参照文献[9]的方法,对第一组和第二组数据分别筛选出 2000 个和 1994 个差异表达基因构成完全矩阵;第三组是人类 Hela 细胞周期数据[10](Data3),包含 29 621 个基因和 114 个样本;参照文献[11]的方法筛选出 1100 个与细胞周期最相关的基因构成完全矩阵。在这三组实验数据中,第一组和第二组数据是非时序数据;第三组数据是时序数据。

为了评估缺失值估计算法的性能,首先,对数据进行对数化处理,并将完全矩阵按列进行单位化处理,再从完全矩阵中按比例随机移除元素,得到包含缺失值的表达矩阵。然后,在缺失比例为 1%、2%、5%、10% 和 20% 的情况下,分别使用 KNN 法、SVR 法和 KNN-SVR 法对 Data1、Data2 和 Data3 中的缺失数据进行估计。在 KNN 法中确定 K 值为 15;在 KNN-SVR 法中,首先设定 K 值为 20,以保证支持向量的个数在 15 左右,使 KNN 法和 KNN-SVR 法具有可比性;同时,还将 K 设定为不同的值,对缺失值进行估计。在 SVR 法中使用径向基核函数。在参数优化过程中,采用网格搜索(grid search)策略,对训练集进行 5 重交叉验证,其中,参数 C 和 σ^2 的范围设置为 $[2^3, 2^5]$,步长为 2,ε 为 0.01。最后,在每个缺失百分比和 K 值情况下,每种方法重复实验 50 次,取 NRMSE 的均值和方差作为性能评估指标,其中 NRMSE 如下所示:

$$\mathrm{NRMSE}=\sqrt{\frac{\mathrm{mean}[(f(x)-y)^2]}{\mathrm{E}(y)}} \tag{6}$$

其中,$f(x)$ 为估计值,y 为原始值,$\mathrm{E}(y)$ 是 y 的方差。

算法的程序均在 C++ 平台上实现。

3.1 NRMSE 均值对比

在不同缺失比例的情况下,KNN 法、SVR 法和 KNN-SVR 法在 Data1、Data2 和 Data3 上的 NRMSE 均值如图 1 所示。NRMSE 均值越大,估计精度越低。KNN 和 SVR 分别表示 KNN 法和 SVR 法的性能评估曲线,KS20/100/200/400/800/1200 分别表示 KNN-SVR 法在 K 值为 20/100/200/400/800/1200 时的性能评估曲

线。

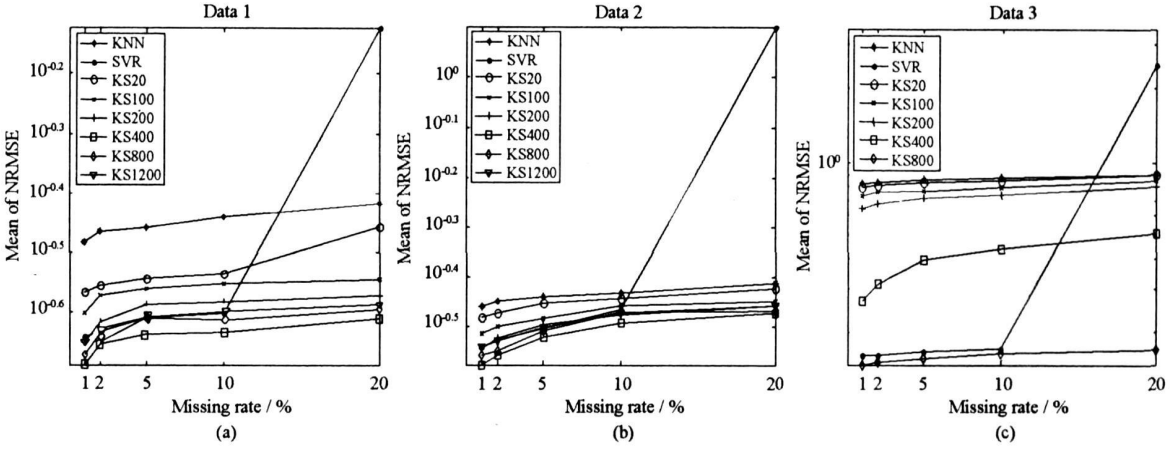

图 1 在不同缺失比例的情况下, 几种方法在三个数据集上的 NRMSE 均值比较

Fig. 1 Means of NRMSE on Data1/2/3 estimated with the three methods under different missing rates

从图 1 中可以看出, KNN 法效果最差; 与非时序数据相比, KNN 法在时序数据中的精度非常低, 表明KNN 法不适用于时序的数据; 而KS20 略优于KNN 法是由于在 KS20 中的 SVR 利用了基因间的相关信息。KS20/100/200 中训练集基因过少, 大量信息流失, 效果次于SVR 法; 但当缺失比例为 20% 时, SVR 法的精度很低, 表明 SVR 法不适用于缺失率高的数据。

在图 1(a) 和图 1(b) 中, 随着 K 值的增加, KNN-SVR 法的性能逐渐升高, 当 K 值达到 400 时性能最佳, 随后开始下降。这是由于随着 K 值的增加, 训练集中相关基因越来越多, 估计精度就越来越高; 但是当 K 值增加到一定程度时, 训练集中加入了许多相似度较低的基因, 所以精度就有所降低。在图 1(c) 中, KNN-SVR 法的精度随着 K 值的增加而不断提高, 当 K 到达 800 时才得到优于 SVR 法的估计精度, 而 K= 800 意味着与全部基因数目 1100 相当; 这是因为 Data3 是细胞周期表达数据, 细胞周期之间存在过渡性, Data3 中基因相似度不高, 所以在数据较多时才能挖掘出比较可靠的数据间的关联性信息。

3.2 NRMSE 方差对比

多次实验结果的方差可以反映缺失值估计方法的稳定性, 同时也可反映估计值的可信度。方差越小, 稳定性越高, 说明可信度越高。为了比较稳定性, 本文在缺失比例为 5% 时, 对各方法在三组数据上的NRMSE 方差进行了统计, 如图 2 所示。

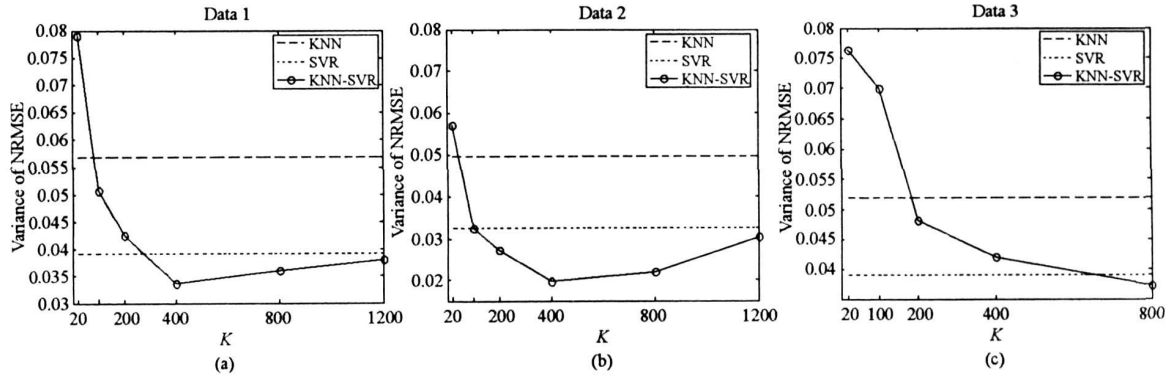

图 2 在缺失比为 5% 时, 几种方法在三个数据集上的 NRMSE 方差比较

Fig. 2 Variances of NRMSE on Data1/2/3 estimated with the three methods under 5% missing rate

从图 2(a) 和图 2(b) 中可以看出, 当 K 值较小时, KNN-SVR 法的方差比 KNN 法和 SVR 法的都要大; 随着 K 值的增加, 方差越来越小, 当 K 值到达一定程度时, KNN-SVR 法的方差要比 KNN 法和 SVR 法的小得多; 当 K 值再增加时, KNN-SVR 法的方差又有所上升, 与 SVR 法的相当。在图 2(c) 中, 随着 K 值的

增加, 方差一直减小, 直到 K 值与全体基因数目相当时, KNN-SVR 法的方差才与 SVR 法的相当。原因与前一部分基本相同。还需要补充的一点是, 在 KNN-SVR 法和 SVR 法中, 当训练集中基因数目过小或过大时, 由于可利用的信息少或加入的干扰太多, 都使得优化过程中的波动比较大, 所以方差较大。

综上所述, 可以得到如下结论: KNN 法不适用于时序数据的缺失估计; SVR 法不适合处理缺失率较高的数据; KNN-SVR 法无论用于时序数据还是用于非时序数据, 都可以得到较高的精度, 尤其在缺失率较高时, 可以得到很好的结果, 而且 KNN-SVR 法的稳定性要明显优于上述两种方法; 由于 KNN-SVR 法要在选出的 K 个近邻中筛选出支持向量, 当 K 很小时 SVR 的精确度会比较低, 所以 K 的取值一般比较大, 而且相对基因总数来说, 在非时序数据中较小, 在时序数据中较大。

4 结 论

有效的缺失值估计能够提高基因表达谱后续分析的准确性和稳定性。本文提出的基于 KNN-SVR 的缺失值的估计方法, 利用与目标基因具有较高相似度的基因子集对缺失值进行估计, 消除了不相似基因的干扰。实验结果表明, 基于 KNN-SVR 的缺失值的估计方法不但提高了缺失估计的精度, 而且具有较高的稳定性。该方法为缺失值的有效处理提供了一种新思路, 有助于基因表达谱的后续分析得到更加准确的结果。

参 考 文 献:

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

[2] Olga T, Michael C, Sherlock G, et al. Missing Value Estimation Methods for DNA Microarrays[J]. Bioinformatics, 2001, 17: 520- 525.

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

[4] Kim H, 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.

[5] Wang X, Li A, Jiang Z H, et al. Missing Value Estimation for DNA Microarray Gene Expression Data by Support Vector Regression Imputation and Orthogonal Coding Scheme[J]. BMC Bioinformatics, 2006, 7:32.

[6] 杜树新, 吴铁军. 用于回归估计的支持向量机方法[J]. 系统仿真学报, 2003, 15(11): 1580- 1585.

[7] Alon U, Barkai N, Notterman D A, et al. Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays[J]. Proc. Natl. Acad. Sci., USA, 1999, 96(6): 6745- 6750.

[8] Golub T R, Slonim D K, Tamayo P, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring [J]. Science, 1999, 286(15): 531- 537.

[9] Nathalie P, Frank D S, Johan A K, et al. Systematic Benchmarking of Microarray Data Classification: Assessing the Role of Non-linearity and Dimensionality Reduction[J]. Bioinformatics, 2004, 20(17): 3185- 3195.

[10] http://genome-www.stanford.edu/Human- CellCycle/HeLa/.

[11] Whitfield M L, Sherlock G, Saldanha A J, et al. Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors[J]. Molecular Biology of the Cell, 2002, 13: 1977- 2000.