基于精简关联度的基因表达数据迭代填补算法

何 云 皮德常
(南京航空航天大学计算机科学与技术学院 南京 210016)

摘 要 基因表达数据时常出现缺失,阻碍了对基因表达的研究。提出了一种新的相似性度量方案——精简关联度,在此基础上,又提出了基于精简关联度的缺失数据迭代填补算法(RKNNimpute)。精简关联度是对灰色关联度的一种改进,能达到与灰色关联度同样的效果,却显著降低了算法的时间复杂度。RKNNimpute算法以精简关联度作为相似度量,将填补后的基因扩充到近邻的候选基因集,通过迭代的方式填补其他缺失数据,提高了算法的填补效果和性能。选用时序、非时序、混合等不同类型的基因表达数据集进行了大量实验来评估RKNNimpute算法的性能。实验结果表明,精简关联度是一种高效的距离度量方法,所提出的RKNNimpute算法优于常规填补算法。

关键词 基因表达数据,精简关联度,填补,迭代,缺失值
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.051

Iterative Imputation Algorithm Based on Reduced Relational Grade for Gene Expression Data

HE Yun PI De-chang
(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract Gene expression data frequently suffers from missing value, which adversely affects downstream analysis. A new similarity metric method named reduced relational grade was proposed. Based on this, we presented the iterative imputation algorithm for gene expression data (RKNNimpute). Reduced relational grade is an improvement of gray relational grade. The former can achieve the same performance as the latter while greatly reducing the time complexity. RKNNimpute imputes missing value iteratively by considering the reduced relational grade as similarity metric and expanding the set of candidate genes to nearest neighbors with imputed genes, which improves the effect and performance of the imputation algorithm. We selected data sets of different kind, such as time series, non-time series and mixed, and then experimentally evaluated the proposed method. The results demonstrate that the reduced relational grade is effective and RKNNimpute outperforms common imputation algorithms.

Keywords Gene expression data, Reduced relational grade, Imputation, Iteration, Missing value

1 引言

分析和利用基因表达数据已成为人类基因组计划的重要内容。基因表达数据表示基因转录产物mRNA在细胞中的丰度,蕴含着基因功能和基因表达调控信息,反映了细胞当前的生理状态。基因表达数据的分析研究在医学临床诊断、药物疗效判断、疾病发生的机理等方面都有应用。

DNA微阵列技术是检测基因表达水平的常用方法[1],它可以在一块基因芯片上,同时检测到不同样本的成千上万条基因在不同组织状态中mRNA的表达水平。它在一块带有DNA微阵列涂层的特殊玻璃片上,固定大量的核酸探针分子,然后与荧光标记的样本中的DNA、cDNA、RNA等进行杂交,检测探针分子杂交信号的强弱来获取样本基因序列信息。

微阵列技术有快速、精确、低成本的优点,但在实验过程中,总有内部或外部因素导致某些数据无法获取,最终获得的基因表达数据含有缺失值。有的包含10%的缺失值,更有甚者,在某些数据集中高达90%的基因含有一个或者多个缺失值[2]。基因表达数据产生缺失的常见原因有:基因芯片表面有灰尘或者划痕、基因分解不完全、图像损坏、杂交失败、实验方法错误等[3]。

现在很多数据分析方法只能处理完整的数据集,如层次聚类、主成分分析(PCA)、奇异值分解(SVD)等。缺失数据妨碍了数据的分析和规律的获取,必须采取合适的方法进行处理。重复实验不失为解决基因表达数据缺失的一种办法,但是其实验过程复杂、实验费用高昂、耗时很长,在实际微阵列实验中不可取。若缺失率很小或者为了处理方便,在某些数据处理中可直接删除含缺失值的数据项。基因表达数据是多元数据,这无疑会删除大量有用的信息,导致最终基因分析结果不可靠[4]。缺失值填补技术成本低,无需重复实验就能高效恢复缺失数据,是处理数据缺失问题的有效方法[5]。对于少量能容忍缺失值存在的机器学习算法,事先的填补也可以明显提高算法的性能。文献[6]在软件项目数据上进行实验,

到稿日期:2014-11-24 返修日期:2015-03-25 本文受国家自然科学基金(U1433116),江苏省"333"高层次人才工程,航空科学基金(20145752033)资助。

何 云(1991—),女,硕士生,主要研究方向为数据挖掘,E-mail:1561850387@qq.com;皮德常(1971—),男,博士,教授,博士生导师,CCF会员,主要研究方向为数据挖掘、大数据管理与分析。

证实了 K 近邻填补算法能显著提高可容忍缺失值的 C4.5 算法的性能。

本文针对基因表达数据的缺失现象，提出了一种基于新的关联度的迭代填补算法 RKNNimpute。该算法是对 K 近邻算法的改进，以精简关联度作为距离度量，大幅度降低了时间复杂度。同时，RKNNimpute 利用了不完整（含缺失值）基因的信息，对数据利用率高，并且运用融入迭代，能对缺失数据进行迭代填补。本文选取不同类型的基因表达数据集，进行了大量实验，结果表明，在计算基因间相似性时，精简关联度能达到与灰色关联度一样优秀的效果，并能显著地降低时间复杂度。在实际的基因表达数据集中，RKNNimpute 算法比现有填补算法性能更优。

本文第 2 节分析了数据填补中的常规方法；第 3 节介绍了灰色关联度的概念；第 4 节提出了精简关联度的思想，并给出了基于精简关联度的迭代填补算法的实现过程；第 5 节通过大量实验分别评估了精简关联度和 RKNNimpute 算法的性能；最后对全文进行了总结。

2 相关工作

K 近邻填补算法（KNNimpute）为每一个含缺失值的基因找出与之距离最近的 K 个完整的邻居基因，根据邻居基因在缺失属性上的值，使用加权平均来填补缺失的基因数据。K 近邻填补算法的思想是，数据中同一个类别或者性质相似的事物有很多特征是相同的，对于同一个基因表达数据集，相似的基因在相似的样本实验上会有相似的表现。基于这样的事实，KNNimpute 实现了基因表达缺失数据的填补。

文献[7]针对不同缺失率的基因表达数据，将基于奇异值分解的填补算法、K 近邻填补算法和均值填补法进行实验分析比较，结果表明 KNNimpute 有更好的健壮性和准确率。文献[8]综合阐述了 23 种填补算法对基因表达数据的预测性能，发现 KNNimpute 填补准确率高。大量文献对 KNNimpute 进行了研究，并提出了一些改进算法，改进主要是针对两个方面，即填补的顺序和距离度量。其中，SKNNimpute[9] 对数据缺失率进行了排序，按缺失率由低到高的顺序完成填补工作，并将填补后的基因扩充到近邻的选择集中。壳近邻填补法[10] 仅选用缺失属性的左右近邻预测缺失值，以交叉验证法确定最近邻的数目。文献[11]提出了基于马氏距离的 K 近邻填补算法，马氏距离能够消除变量间相关性的影响，但计算过程比较繁琐。目前流行的 KNNimpute 是根据欧氏距离来选择邻居基因，欧氏距离只是将两个基因在不同属性上的差异进行简单叠加，没有综合考虑整个数据集。文献[12]表明用灰色关联度来计算两个样本之间的相似度比欧氏距离或其他距离度量更合适。

现存的很多填补算法都是单一值填补，很难提供有效的标准误差和置信区间，这是因为它们忽视了数据集本身的不确定性，低估了方差[12]。文献[13]提出了无参的迭代填补法，能应对缺乏数据分布的先验知识的情况，并实验证实迭代比普通的单一值填补准确率高。文献[12]提出了基于灰色关联度的 K 近邻迭代填补算法（GKNNimpute），以灰色关联度为相似度量，选取目标基因的 K 个近邻基因，加权平均预测缺失值，当数据集中所有缺失值都被填补后，再进行新一轮的填补，直到算法收敛才停止迭代。GKNNimpute 对连续型数据和离散型数据都能取得较好的填补效果，但是它更适合小型数据集，当数据集的规模进一步扩大时，计算灰色关联度将耗费大量时间。本文在 GKNNimpute 算法的基础上，提出了 RKNNimpute 算法，以精简关联度作为距离度量，可显著降低时间复杂度，且保持算法的良好填补性能。

3 灰色关联度

1982 年邓聚龙教授提出了灰色系统理论[14]，它擅长分析复杂系统并得到可靠结果，目前已经得到广泛应用。灰色关联分析用于寻求各子系统之间的数值关系，以衡量因素之间的相似程度。灰色关联分析可以将两个系统的发展趋势进行量化，若两个系统同步化程度高、变化趋势一致，那么它们的关联程度较高，反之较低。

假定基因表达数据集 X 是 $n\times m$ 的矩阵（n 远大于 m），$X^{complete}$ 表示 X 中不含缺失值的基因构成的完整数据矩阵（不含缺失数据）。待填补的基因记作目标基因，为目标基因 x_i 提供有用的预测信息的基因称为候选基因。其中 x_i 表示基因 $x_i=[x_{i1},x_{i2},\cdots,x_{im}]$，$x_{ij}$ 表示基因 x_i 在实验环境 j 下的表达值。

灰色关联系数（Gray Relational Coefficient，GRC）可用来描述目标基因与候选基因在特定属性列上的相似程度，用 x_i 表示目标基因，x_j 表示候选基因。灰色关联系数的定义如下：

$$GRC(x_{iq},x_{jq})=\frac{\min_{\forall t}\min_{\forall k}|x_{ik}-x_{tk}|+\rho\max_{\forall t}\max_{\forall k}|x_{ik}-x_{tk}|}{|x_{iq}-x_{jq}|+\rho\max_{\forall t}\max_{\forall k}|x_{ik}-x_{tk}|} \tag{1}$$

其中，ρ 是分辨系数，一般在 0～1 之间取值，通常取 $\rho=0.5$。

灰色关联度（Gray Relational Grade，GRG）是将各属性下的灰色关联系数进行综合，得到一个数值，作为两个基因之间相似程度的度量。目标基因 x_i 与候选基因 x_j 之间的灰色关联度为：

$$GRG(x_i,x_j)=\frac{1}{m}\sum_{k=1}^{m}GRC(x_{ik},x_{jk}) \tag{2}$$

灰色关联度从全局的角度衡量两个基因间的相似度，克服了欧氏距离的不足。在选取最近邻的过程中，灰色关联度值越大，表明基因之间相似性越大，差异越小；反之说明基因之间的相似性越小，差异越大。

4 RKNNimpute 算法

4.1 改进的关联度

高维的基因表达数据集描述了成千上万条基因在不同实验环境下的表达情况。从式(1)可以发现，对于每一个目标基因，在计算灰色关联系数时，必须检索全局数据集中的每一个候选基因，这对于基因表达数据而言，时间开销非常大。在式(1)的基础上，本文提出了一种新的关联系数计算方法，它是对灰色关联系数的一种简化，记作精简关联系数（Reduced Relational Coefficient，RRC）。精简关联系数可用来衡量目标基因 x_i 和候选基因 x_j 在特定属性上的相似度。精简关联系数的定义如下：

$$RRC(x_{iq},x_{jq})=\frac{\rho_{Max\ \forall k}\{|x_{ik}-\min_k|,|x_{ik}-\max_k|\}}{|x_{iq}-x_{jq}|+\rho_{Max\ \forall k}\{|x_{ik}-\min_k|,|x_{ik}-\max_k|\}} \tag{3}$$

函数 $Max\{,\}$ 表示两者中的较大值；$\min_k$ 和 $\max_k$ 分别指

代属性 k 上的最小值和最大值，可在数据读取或者数据归一化的过程中求得，在计算精简关联度时直接当作已知量。

基因表达数据集的数据间具有很强的相关性，经大量实验发现，$\min_{\forall t}\min_{\forall k}|x_{ik}-x_{tk}|$ 的数值一般趋向于数值 0，同时公式 $\max_{\forall t}\max_{\forall k}|x_{ik}-x_{tk}|$ 的计算结果总是趋向于数值 $Max_{\forall k}\{|x_{ik}-\min_k|,|x_{ik}-\max_k|\}$。式(3)是对式(1)的简化，将式(1)中占用大量计算量的 $\min_{\forall t}\min_{\forall k}|x_{ik}-x_{tk}|$ 和 $\max_{\forall t}\max_{\forall k}|x_{ik}-x_{tk}|$ 的计算结果直接用相应的极限值 0 和 $Max_{\forall k}\{|x_{ik}-\min_k|,|x_{ik}-\max_k|\}$ 进行近似代替，极大地降低了时间复杂度。

$RRC(x_{iq},x_{jq})$ 在 0～1 之间取值，$RRC(x_{iq},x_{jq})$ 的值越大，表明 x_{iq} 与 x_{jq} 之间越接近。当 $x_{iq}=x_{jq}$ 时，$RRC(x_{iq},x_{jq})=1$ 达到最大值；当 x_{iq} 与 x_{jq} 相差很大时，$RRC(x_{iq},x_{jq})$ 达到最小值。目标基因 x_i 与候选基因 x_j 之间的精简关联度(Reduced Relational Grade，RRG)为：

$$RRG(x_i,x_j)=\frac{1}{m}\sum_{k=1}^{m}RRC(x_{ik},x_{jk}) \tag{4}$$

相应地，精简关联度也是每个属性上相似性的综合，精简关联度的值越大，基因间的相似性越大。

假设候选基因集的大小为 $N*m$，对于任一目标基因，计算它与候选基因集中所有基因的灰色关联度，时间复杂度是 $O(N*N*m)$；而计算精简关联度时，时间复杂度简化为 $O(N*m)$。对于基因表达数据集，$N\gg m$，因而精简关联度在时间复杂度上的提高非常大。

4.2 填补缺失值

给定目标基因 x_i(属性 t 上的值缺失)，计算它与所有候选基因的精简关联度，选取与之最相近的 K 个邻居基因，对相应缺失属性上的值进行加权平均，填补基因的缺失值。

$$\hat{y}_{it}=\sum_{k=1}^{K}w_{ik}x_{kt} \tag{5}$$

w_{ik} 表示第 k 个邻居基因 x_k 与目标基因 x_i 之间的权值。

$$x_{ik}=RRG(x_i,x_k)/\sum_{k=1}^{K}RRG(x_i,x_k) \tag{6}$$

4.3 数据归一化

归一化可消除量纲，缩小不同值之间的差异，避免数据偏置。本文采用最小-最大规范法，将数据归一化在 0～1 之间。假设 $\max(t)$ 和 $\min(t)$ 分别为属性 t 上的最大值和最小值，表示基因 x_i 在属性 t 上的数值，归一化公式为：

$$\tilde{x}_{it}=\frac{x_{it}-\min(t)}{\max(t)-\min(t)} \tag{7}$$

4.4 RKNNimpute 算法

RKNNimpute 通过计算目标基因与候选数据集中所有基因的精简关联度 RRG，选取与之最相近的 K 个邻居，采用加权平均法，填补目标基因中的缺失值。算法采用迭代方法，对整个数据集进行多次填补，直到满足条件为止。

算法 1 RKNNimpute

Input：含缺失值的基因表达数据集 X

Output：完整的基因表达数据集

1. Step1：Initialization
2. FOR each target gene x_i in X
3. 用行(基因)均值填补 x_i 中的缺失值
4. END FOR//获得完整数据集 $X^{complete}$
5. 对 $X^{complete}$ 进行归一化，得到 $X^{complete(0)}$
6. Step2：Imputation
7. h=0
8. REPEAT
9. h++//第 h 次迭代填补(h=1,2,3,…)
10. FOR each target gene x_i in X
11. 在 $X^{complete(h-1)}$ 的基础上构造候选基因矩阵
12. FOR each candidate gene x_j
13. 计算 $RRG(x_i,x_j)$
14. END FOR
15. 选择精简关联度最大的 K 个近邻基因
16. 根据式(6)填补 x_i 中的缺失值
17. END FOR//获得了 $X^{complete(h)}$
18. FOR 每一个填补值 $\hat{y}_j^{(h)}\in X^{complete(h)}$，$\hat{y}_j^{(h-1)}\in X^{complete(h-1)}$
19. 计算 $\delta^{(h)}=\frac{1}{N}\sum_{j=1}^{N}(\hat{y}_j^{(h)}-\hat{y}_j^{(h-1)})^2$
20. END FOR
21. UNTIL($\delta^{(h)}<\tau$ or 达到最大迭代次数)

本文选用均方根误差(Root Mean Squared Error，RMSE)来衡量算法填补的准确性：

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i-y_i)^2}{N}} \tag{8}$$

其中，y_i 表示准确值也就是真实值；$\hat{y}_i$ 表示预测估计值；N 表示缺失数据的个数。$RMSE$ 越小，表明估计值与真实值之间越接近，算法的填补准确率越高。

5 实验与分析

5.1 基因表达数据集

基因表达数据一般分为时序、非时序和混合类型[15,16]，微阵列技术获取的基因表达数据集常常以矩阵的形式表示。矩阵的行表示基因，列(属性)表示样本(实验环境，如生物分子扫描、杂交的方式，细胞周期的不同阶段，肿瘤的类型等)。本文根据不同类型的数据集、不同观测对象和不同的实验点，选取了有代表性的 3 个不同数据基因表达数据集，它们均来自国际公共基因数据库：http://www.ncbi.nlm.nih.gov/geo/。

数据集一(transcription)属时序数据，为厌氧环境下生长的酿酒酵母被暴露在空气中的转录反应，有 6 个属性，依次表示酿酒酵母在空气中暴露 0 分钟、5 分钟、10 分钟、20 分钟、60 分钟和 120 分钟。

数据集二(mutant)属非时序数据，是热敏感型的 mex67-5 突变体和野生型的 Mex67 菌株在 37℃下 RNA 的比较，有 6 个属性，表示培养了 6 个样本，进行了 6 次比较实验。

数据集三(salmonella)属混合数据，是肠道沙门氏菌被过氧化氢溶液处理之后的基因表达情况，有 6 个属性，前 3 个属性表示沙门氏菌在过氧化氢溶液中处理了 2 个小时，后 3 个属性表示在过氧化氢溶液中处理了 3 个小时。

数据集 transcription，mutant，salmonella 分别属于时序、非时序和混合类型 3 种不同类型的数据集，都存在不同程度的数据缺失现象，其中数据集一的缺失情况尤为严重。为了衡量填补算法的性能，在实验之前进行预处理，剔除数据集中含缺失数据的基因，得到 3 个完整的数据集。预处理前后，基因表达数据集的信息如表 1 所列。

表 1　基因表达数据集信息

数据集名称	原始数据集		完整数据集		数据缺失率	数据集类型
	行数	列数	行数	列数		
transcription	6495	6	3491	6	12.07%	时序
mutant	7684	6	7106	6	1.46%	非时序
salmonella	5184	6	5151	6	0.11%	混合类型

为了表示方便，完整数据集 transcription，mutant 和 salmonella 分别记作数据集 a，b，c。

5.2　RKNNimpute 算法性能分析

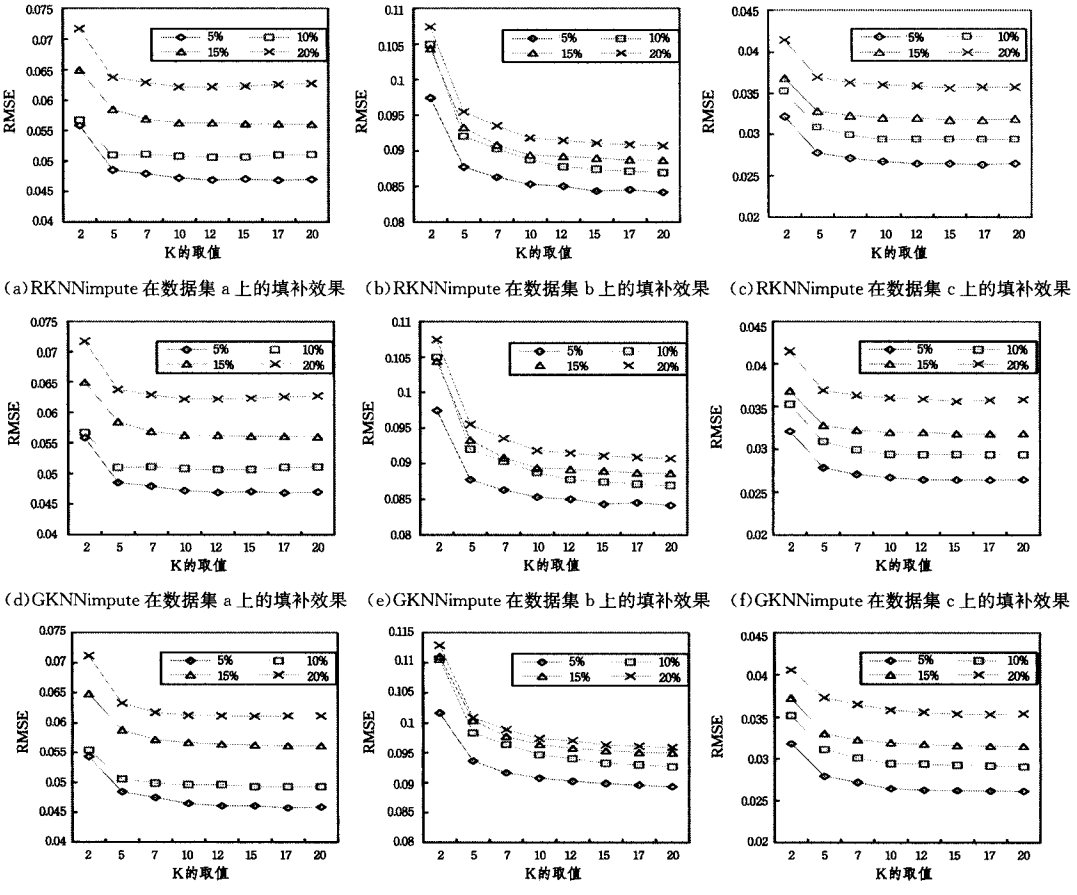
为了分析本文提出的精简关联度以及 RKNNimpute 算法的性能，选用基于灰色关联度的迭代算法[12]（GKNNimpute）和迭代 K 近邻填补算法[15]（IKNNimpute）（在普通的 K 近邻算法基础上添加了迭代过程）作为参考，从填补准确率和时间复杂度两方面进行实验分析比较。

在预处理后得到的完整数据集 a，b，c 的基础上，每次实验前引进不同比率的随机缺失值，用填补算法处理随机得到的不完整的数据集，采用均方根误差来衡量算法的准确率。

5.2.1　参数设置

K 近邻方法及其改进算法都有一个共同点：算法的性能依赖于 K 的取值。经实验发现，K 的值在一定程度上与数据的类型和缺失比率的大小有关，但在理论上，并没有一个确切的公式。K 值过大，则近邻基因过多，相对于目标基因而言，相似性会显得不足，同时，噪声数据对预测的消极影响也会变大，从而降低了预测准确率；K 值过小，强化了某几个近邻基因，造成预测结果的误差偏大。文献[15]针对 K 近邻算法设计出自动获取最佳 K 值的程序，发现当 K 在 10～20 之间取值时，K 值的变化对预测准确率的影响极小。

为了选取合适的 K 值，进行实验来评估当 $K=2,5,7,10,12,15,17,20$ 时算法填补性能受 K 取值的影响。设置收敛精度 τ 为 0.01，最大迭代次数为 10。实验结果如图 1 所示。其中，图 1(a)－图 1(c)展示了取不同的 K 值时算法 RKNNimpute 在数据集 a，b，c 上的填补情况。相应地，图 1(d)－图 1(f)和图 1(g)－图 1(i)分别描述了 K 值对算法 GKNNimpute 和算法 IKNNimpute 的影响。



(a)RKNNimpute 在数据集 a 上的填补效果　(b)RKNNimpute 在数据集 b 上的填补效果　(c)RKNNimpute 在数据集 c 上的填补效果

(d)GKNNimpute 在数据集 a 上的填补效果　(e)GKNNimpute 在数据集 b 上的填补效果　(f)GKNNimpute 在数据集 c 上的填补效果

(g)IKNNimpute 在数据集 a 上的填补效果　(h)IKNNimpute 在数据集 b 上的填补效果　(i)IKNNimpute 在数据集 c 上的填补效果

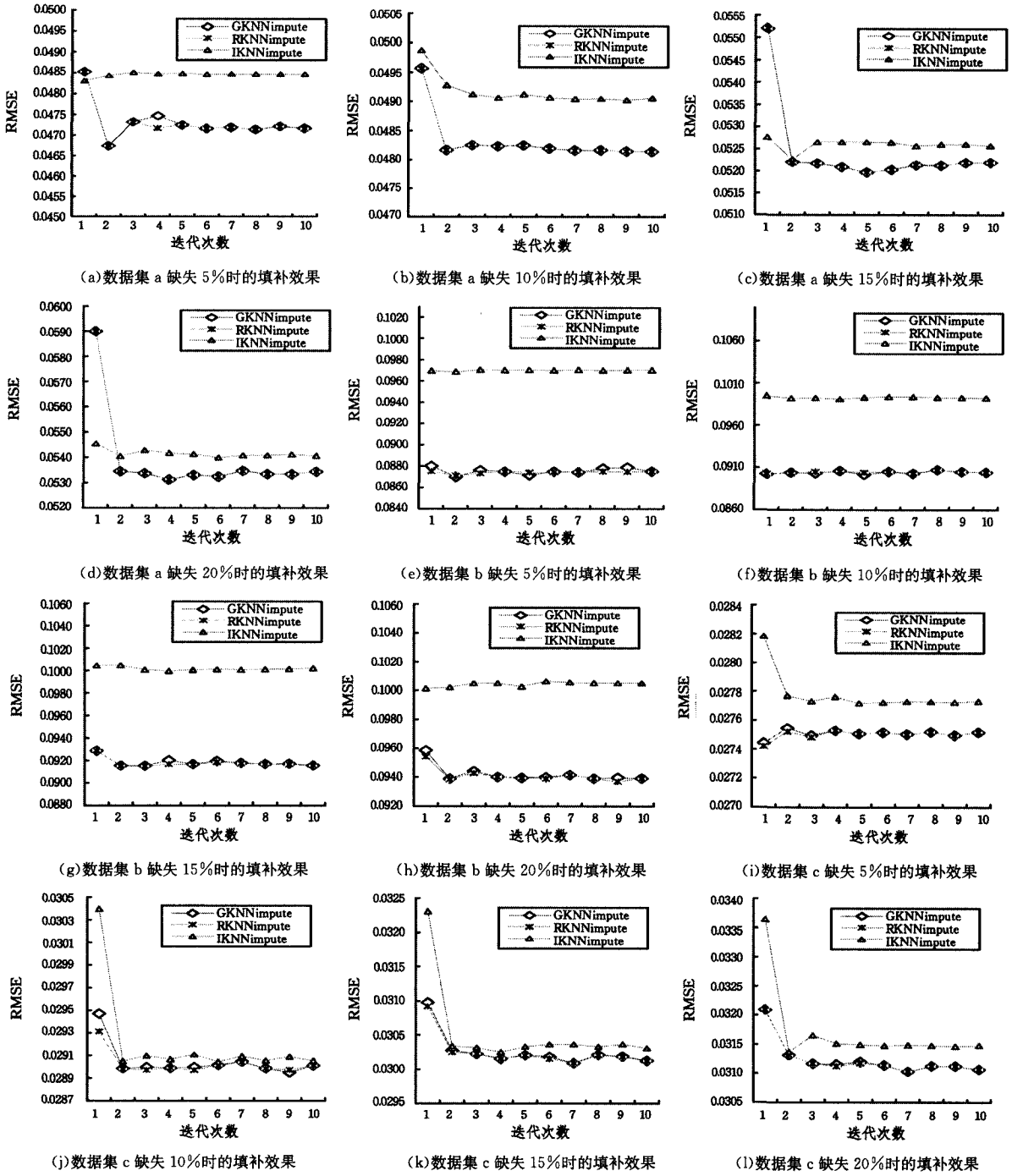
图 1　K 值的变化对算法填补性能的影响

事实上，随着数据缺失率的提高，3 个算法的填补误差都显著增大，即算法填补的准确率很大程度上受到缺失率的影响；当 K 在 10～17 之间取值时，K 值的波动对算法填补性能的影响非常小，此时算法的填补误差处于一个较低的水平。因此，接下来将取 $K=10$ 进一步进行实验。

5.2.2　填补性能

从填补准确率和时间复杂度两个方面对 RKNNimpute 算法与 GKNNimpute 和 IKNNimpute 进行实验比较和分析。

RKNNimpute、GKNNimpute 和 IKNNimpute 都是迭代类型的算法，前一次迭代得到的填补值（预测值）将作为下一次迭代过程中缺失值的替代值，在此基础上对数据集进行新一轮的填补。迭代是一步步改善填补值的过程。下面将具体展示 3 个算法在进行一次填补实验过程中每一次迭代填补的情况，取迭代次数为 10，如图 2 所示。



(a)数据集 a 缺失 5%时的填补效果　(b)数据集 a 缺失 10%时的填补效果　(c)数据集 a 缺失 15%时的填补效果

(d)数据集 a 缺失 20%时的填补效果　(e)数据集 b 缺失 5%时的填补效果　(f)数据集 b 缺失 10%时的填补效果

(g)数据集 b 缺失 15%时的填补效果　(h)数据集 b 缺失 20%时的填补效果　(i)数据集 c 缺失 5%时的填补效果

(j)数据集 c 缺失 10%时的填补效果　(k)数据集 c 缺失 15%时的填补效果　(l)数据集 c 缺失 20%时的填补效果

图 2　算法 RKNNimpute、GKNNimpute 和 IKNNimpute 在数据集 a,b,c 上的填补效果比较

表 2　算法耗时(ms)

	数据集 a				数据集 b				数据集 c			
缺失率	5%	10%	15%	20%	5%	10%	15%	20%	5%	10%	15%	20%
GKNNimpute	2978	5388	6876	8096	11582	21032	28565	31979	6185	10925	13945	18185
RKNNimpute	2755	4888	6443	7485	12406	22693	30418	39362	6191	10662	14620	17512
IKNNimpute	4111439	7176826	8904707	8904770	33457400	60015902	79448905	96778006	14534205	2330622	29330607	36188027

其中,图 2(a)－图 2(d)是 3 个填补算法在时序数据集 a 上迭代 10 次的填补效果,数据集 a 依次标记 5%,10%,15%,20%的随机缺失。同理,图 2(e)－图 2(h)和图 2(i)－图 2(l)分别是算法在非时序数据集 b 和混合类型数据 c 上的填补效果。

首先,对 RKNNimpute 和 GKNNimpute 的填补准确率进行比较,从图 2 中可观察到两个算法的预测曲线基本重合,每一次迭代的填补准确率基本相同,说明算法 RKNNimpute 和 GKNNimpute 有近似相同的填补准确率。

对 RKNNimpute 和 IKNNimpute 的填补准确率进行比较,从图 2 中可观察到 RKNNimpute 算法的误差更小,预测准确率明显优于 IKNNimpute 算法。

表 2 描述了进行上述对比实验时 3 个算法的耗时情况,

(下转第 283 页)

of heaviside function[J]. Journal of Harbin University of Science and Technology,1987,45(4):760-782

[16] 李慧玲,张春阳,李卓识,等.解非凸优化问题的一个同伦内点方法[J].东北师大学报(自然科学版),2009,41(4):764-785
Li Hui-ling,Zhang Chun-yang,Li Zuo-shi,et al. Homotopy interior point method for non-convex optimization problem with weak pseudo cone condition[J]. Journal of Northeast Normal University(Natural Science Edition),2009,41(4):764-785

[17] Gupta S. A note on the asymptotic distribution of lasso estimator for correlated data[J]. Sankhya A,2012,74(1):10-28

[18] Kutyniok G. Theory and applications of compressed sensing[J]. Gamm-mitteilungen,2013,36(1):79-101

[19] Gallavotti G. Chaotic hypothesis:onsager reciprocity and fluctuation-dissipation theorem[J]. Journal of Statistical Physics,1996,84(5/6):899-925

[20] 王定成,苏淳. b值独立同分布随机变元序列的矩完全收敛性[J].应用数学学报,2004,27(3):440-448
Wang Ding-cheng, Sun Chun. Moment Complete Convergence for Sums of i. i. d. Random Elements in Banach Space[J]. Acta Mathematicae Applicatae Sinica,2004,27(3):440-448

(上接第255页)

可以发现GKNNimpute算法耗时最多,RKNNimpute和IKNNimpute算法相对快捷。

综上,RKNNimpute与GKNNimpute相比,二者有近似相等的填补准确率,RKNNimpute具有更小的计算复杂度。通过比较基于精简关联度迭代算法(RKNNimpute)和基于灰色关联度迭代算法(GKNNimpute)的填补性能(两算法仅相似性的度量方法不同,其他细节均相同),间接比较精简关联度(RRG)和灰色关联度(GRG)的性能。在相似性衡量上,精简关联度能达到与灰色关联度完全相同的效果;在时间开销上,精简关联度大大降低了算法的时间复杂度。因而,本文提出的精简关联度是一种优秀的距离度量方法。与IKNNimpute算法比较可知,RKNNimpute算法准确率更高,针对基因表达数据能取得不错的填补效果。

结束语 本文提出了新的关联度计算方案——精简关联度,它是对灰色关联度的改进。选用3种类型的数据集,设置不同的缺失率,进行大量实验,分析了精简关联度的性能。在预测准确性上,它与灰色关联度有一致的相似性度量效果;在时间复杂度上,所提出的度量方法显著地降低了时间复杂度。因而,本文提出的精简关联度是一种高效的相似性度量方法。在此基础上,针对基因表达数据的缺失现象,进一步提出了基于精简关联度的迭代填补算法。RKNNimpute利用填补后的基因和完整基因构成候选基因集,扩大了近邻的选择范围,提高了数据的利用率;通过设置合适的阈值,以迭代的方式提高了填补性能。与迭代K近邻填补算法进行了比较,实验证实RKNNimpute是一种有效的填补算法。

RKNNimpute算法需预先指定迭代终止的条件,即收敛的精度和最大迭代次数,它们能影响算法的迭代次数和填补准确率。如何高效合理地确定收敛精度和最大迭代次数以及进一步提高算法的性能是今后的研究方向。

参考文献

[1] Hoheisel J D. Microarray technology:beyond transcript profiling and genotype analysis [J]. Nature Reviews Genetics, 2006, 7(3):200-210

[2] De Brevern A G, Hazout S, Malpertuy A. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering [J]. BMC Bioinformatics, 2004, 5(1): 114-119

[3] Yang Y H, Buckley M J, Dudoit S, et al. Comparison of methods for image analysis on cDNA microarray data [J]. Journal of Computational and Graphical Statistics, 2002, 11(1): 108-136

[4] Pedro J, Garcia-Laencina, et al. K nearest neighbours with mutual information for simultaneous classification and missing data imputation [J]. Neurocomputing, 2009, 72(7-9): 1483-1493

[5] Moorthy K, Mohamad M S, Deris S. A Review on Missing Value Imputation Algorithms for Microarray Gene Expression Data [J]. Current Bioinformatics, 2014, 9(1): 18-22

[6] Song Qin-bao, Shepperd M, Chen Xiang-ru, et al. Can k-NN imputation improve the performance of C4. 5 with small software project data sets? A comparative evaluation [J]. Journal of Systems and Software, 2008, 81(12): 2361-2370

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

[8] Alan Wee-Chung, Law Ngai-Fong, Yan Hong. Missing value imputation for gene expression data: computational technique to recover missing data from available information [J]. Briefings in Bioinformatics, 2010, 12(5): 498-513

[9] Meng Fan-chi, Cheng Cai, Hong Yan. A Bicluster-Based Bayesian Principal Component Analysis Method for Microarray Missing Value Estimation [J]. Biomedical and Health Informatics, 2014, 18(3): 862-871

[10] Zhang Shi-chao. Shell-neighbor method and its application in missing data [J]. Applied Intelligence, 2011, 35(1): 123-133

[11] 杨涛.基因表达缺失数据填充算法研究[D].长沙:湖南大学,2005
Yang Tao. The research on imputation algorithm of missing values for gene expression data [D]. Changsha: Hunan University, 2005

[12] Zhang Shi-chao. NIIA: Nonparametric Iterative Imputation Algorithm[C] // Trends in Artificial Intelligence, 2008 (PRICAI 2008). Berlin: Springer Berlin Heidelberg, 2008: 544-555

[13] Song Qin-bao, Shepperd M. Predicting software project effort: A grey relational analysis based method [J]. Expert Systems with Applications, 2011, 38(6): 7302-7316

[14] Bras L P, Menezes J C. Improving cluster-based missing value estimation of DNA microarray data [J]. Biomolecular Engineering, 2007, 24(2): 273-282

[15] 李艳芳.基因表达数据的缺失值估计研究[D].哈尔滨:哈尔滨工业大学,2011
Li Yan-fang. Research on missing value imputation for microarray gene expression data [D]. Harbin: Harbin Institute of Technology, 2011