应用概率统计 第 33 卷
第 1 期 2017 年 2 月

Chinese Journal of Applied Probability and Statistics
Feb., 2017, Vol. 33, No. 1, pp. 102-110
doi: 10.3969/j.issn.1001-4268.2017.01.008

基于随机森林模型的成分数据缺失值填补法*

张晓琴* 程誉莹

(山西大学数学科学学院, 太原, 030006)

摘 要: 缺失数据处理是数据挖掘领域中进行数据预处理的一个重要环节, 由于成分数据特殊的几何性质, 传统的缺失值填补方法不能直接用于这种类型的数据. 因此, 对成分数据而言, 缺失值的填补具有十分重要的意义. 为了解决这个问题, 本文利用了成分数据和欧氏数据之间的关系, 提出了一种基于随机森林的成分数据缺失值迭代填补法, 该方法的实施和评估采用模拟和真实的数据集. 实验结果表明: 新的填补方法可广泛应用于多种类型的数据集且具有较高准确性.

关键词: 缺失值填补; 成分数据; 随机森林

中图分类号: O212.1

英文引用格式: Zhang X Q, Cheng Y Y. Imputation of missing values for compositional data based on random forest [J]. Chinese J. Appl. Probab. Statist., 2017, 33(1): 102–110. (in Chinese)

§1. 引 言

在数据分析过程中, 作为机器学习领域基准数据库的UCI数据集中超过40%都含有缺失数据[1]. 其中成分数据是一种重要的具有特殊几何性质的数据类型, 这种数据广泛存在于心理学(时间预算的各种组合)、地质(岩石的矿物成分)、经济学(家庭预算的构图和需求收入弹性), 等等领域. 进行数据挖掘时, 缺失值的处理不当会使相关有价值的信息被忽略[2], 传统的缺失值填补方法直接用于这种类型的数据可能得到不良的结果. Little和Robin[3]从缺失机制将缺失数据划分为完全随机缺失(MCAR)、随机缺失(MAR)和非完全随机缺失(MNAR). 为了更随机的设置缺失位置, 本文选用的缺失类型是完全随机缺失(MCAR), 即缺失数据发生的概率与不完全变量和完全变量都无关.

成分数据的概念可以追溯到1866年Ferrers[4]的工作, D个部分的单形空间定义为

$$S^D = \Big\{ \boldsymbol{x} = [x_1, x_2, \ldots, x_D]^{\mathsf{T}};\, x_i > 0, i = 1, 2, \ldots, D;\, \sum_{i=1}^{D} x_i = c \Big\}. \tag{1}$$

其中, $c > 0$为常数, 常取为1. 如果一个向量$\boldsymbol{x} = [x_1, x_2, \ldots, x_D]^T$属于单形空间$S^D$, 那么$\boldsymbol{x}$叫做一个成分, 维数为$D-1$. 与普通数据相比, 成分数据满足“非负性”和“定和性”, 因此在成分数据上直接利用传统的统计方法会导致一些不合理的结果产生. Aitchison[5]认

*山西省高等学校教学改革项目(批准号: J2014006)、山西省自然科学基金面上项目(批准号: 2015011044)和山西省国际科技合作计划项目(批准号: 2015081020)资助.

*通讯作者, E-mail: zhangxiaoqin@sxu.edu.cn.

本文2016年5月23日收到, 2016年9月24日收到修改稿.

识到成分数据中更关注的是相对信息而非绝对信息, 因此每一个成分变量都对应一个比例. 2000年Aitchison等[6]提出的对数比变换(alr)和中心对数比变换(clr), 以及2003年Egozcue等[7]提出的等距对数比变换(ilr)均可以将一个成分数据向量转换为服从于正态分布的欧氏向量, 经相应的统计推断分析后, 再经过逆变换变回成分数据, 这样有利于传统统计方法的使用. 但是, 大多数统计分析方法是基于完整数据的, 且当数据集中存在缺失值时, 对数比变换将无法实施, 因此成分数据缺失值的处理有很大意义. 本文旨在提出一种填补成分数据缺失值的有效方法.

目前, 关于成分数据缺失值的填补方法主要分为两大类: 参数填补法和非参数填补法. 其中, 参数填补法包括EM算法[8]、迭代回归填补法等[9], 这些算法要求数据维数小于样本量, 变量间不能存在多重共线性, 且单形空间中参数的估计是仍未克服的难题. 非参数填补法包括全局常量填补法(global constant)[10]、属性均值填补法(attribute mean)、乘法替换法[11]、k近邻填补法等, 其中Hron等[9]提出的基于Aitchison距离的k近邻填补法在样本量尽可能多时效果明显, 当样本量较少且缺失率较大时无法找到与缺失样本最接近的样本. 本文将对以上问题进行深入研究. 考虑到随机森林法是以决策树为基础分类器的集成分类器, 在处理欧氏数据过程中, 抗噪能力强, 受异常值影响小, 对数据的分布无限制, 能有效分析高维复杂数据. 本文将对基于随机森林法的成分数据缺失值填补方法进行研究, 期待得到一种更高效的成分数据缺失值填补方法.

本文结构如下: 第2节回顾了一些成分数据的定义和性质. 第3节首先介绍了随机森林填补法, 之后给出了基于成分数据缺失值的随机森林填补方法. 第4节通过模拟数据和实际数据验证了新方法的有效性. 第5节为文章结论.

§2. 成分数据简介

本节将给出成分数据的等距对数比变换及Aitchison距离的定义, 这在后面的实验过程中将会用到.

定义 1 [7]　设向量$\boldsymbol{x}=[x_1,x_2,\ldots,x_D]^{\mathsf{T}}\in S^D$, 令

$$z_i=\sqrt{\frac{D-i}{D-i+1}}\ln\left[\frac{1}{x_i}\sqrt[D-i]{\prod_{l=i+1}^{D}x_l}\right],\qquad i=1,2,\ldots,D-1. \tag{2}$$

公式(2)可将有D个部分的成分向量转化成一个$D-1$维的实向量$\boldsymbol{z}=(z_1,z_2,\ldots,z_{D-1})^{\mathsf{T}}$, 这种变换称为等距对数比变换(isometric log-ratio transformations, ilr). 记作$\boldsymbol{z}=\mathrm{ilr}(\boldsymbol{x})$.

其逆变换为$\mathrm{ilr}^{-1}(\boldsymbol{z})=\boldsymbol{x}=[x_1,x_2,\ldots,x_D]^{\mathsf{T}}$, 即

$$\begin{cases} x_1=\exp\Big\{-\sqrt{\dfrac{D-1}{D}}\,z_1\Big\};\\ x_j=\exp\Big\{\displaystyle\sum_{l=1}^{j-1}\frac{1}{\sqrt{(D-l+1)(D-l)}}\,z_l-\sqrt{\frac{D-j}{D-j+1}}\,z_j\Big\},\qquad j=2,3,\ldots,D-1;\\ x_D=\exp\Big\{\displaystyle\sum_{l=1}^{D-1}\frac{1}{\sqrt{(D-l+1)(D-l)}}z_l\Big\}. \end{cases} \tag{3}$$

经过ilr变换后的数据符合标准的欧氏空间的性质, 并且传统的统计方法也可以使用. 如果一个随机成分向量经过ilr变换后服从多维正态分布: $\mathrm{N}_{D-1}(\mu,\Sigma)$, 则称这个成分向量在单形空间上也服从正态分布. 记作$\boldsymbol{x}\sim\mathrm{N}_S^D(\mu,\Sigma)$, 维数为$D-1$.

定义 2 [6] 设$\boldsymbol{x}=[x_1,x_2,\ldots,x_D]^{\intercal}$, $\boldsymbol{y}=[y_1,y_2,\ldots,y_D]^{\intercal}\in S^D$, 则$\boldsymbol{x}$与$\boldsymbol{y}$的Aitchison距离$d_A$定义为

$$d_A(\boldsymbol{x},\boldsymbol{y})=\sqrt{\Big[\sum_{i=1}^{D}\Big\{\ln\frac{x_i}{g(\boldsymbol{x})}-\ln\frac{y_i}{g(\boldsymbol{y})}\Big\}^2\Big]}=\sqrt{\frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D}\Big(\ln\frac{x_i}{x_j}-\ln\frac{y_i}{y_j}\Big)^2}. \tag{4}$$

且Aitchison距离d_A等于对应的向量经过等距对数比变换后的ilr($\boldsymbol{x}$)和ilr($\boldsymbol{y}$)在欧式空间上的距离d_E, 即

$$d_A(\boldsymbol{x},\boldsymbol{y})=d_E(\mathrm{ilr}(\boldsymbol{x}),\mathrm{ilr}(\boldsymbol{y}))=\|\mathrm{ilr}(\boldsymbol{x})-\mathrm{ilr}(\boldsymbol{y})\|.$$

§3. 基于随机森林的缺失数据填补法

Breiman于2001年首次提出了随机森林法 [12]. 具体过程为: 利用bootstrap重抽样技术, 从原始训练集N中有放回的重复随机抽取b个样本生成新的样本集, 同时生成b个对应的分类树组成随机森林通过这个森林对未知数据进行预测, 选取投票最多的分类. 利用随机森林能广泛处理多种类型的数据及其出色的分类能力, Stekhoven和Bühlmann [13]通过观测值对欧氏数据的缺失值进行预测. 下面首先介绍在欧氏空间上的随机森林缺失值填补法, 然后提出针对成分数据的随机森林缺失值填补法.

3.1 欧氏空间数据缺失值的随机森林填补法 [13]

令$X=(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},\ldots,\boldsymbol{x}^{(p)})$是一个$n\times p$阶的数据矩阵, 对于给定的变量$\boldsymbol{x}^{(s)}$, 其缺失指标集记为$i_{\mathrm{mis}}^{(s)}\subseteq\{1,2,\ldots,n\}$. 首先将数据分成以下四部分: 变量$\boldsymbol{x}^{(s)}$中的观测值记作$y_{\mathrm{obs}}^{(s)}$, 缺失值记作$y_{\mathrm{mis}}^{(s)}$; 其余$p-1$个变量对应的行在$i_{\mathrm{obs}}^{(s)}=\{1,2,\ldots,n\}\backslash i_{\mathrm{mis}}^{(s)}$的数据记作$x_{\mathrm{obs}}^{(s)}$, 对应行在$i_{\mathrm{mis}}^{(s)}$的数据记作$x_{\mathrm{mis}}^{(s)}$.

由于数据缺失的随机性, $x_{\mathrm{obs}}^{(s)}$并非完全已知, $x_{\mathrm{mis}}^{(s)}$也并非完全缺失. 具体填补过程如下:

(i) 利用均值填补或其他简单填补法对X进行初始填补;

(ii) X中缺失的列的指标集记作M, 并将变量(列)按照缺失率由小到大排列;

(iii) 当不满足停止准则γ时:

存储现有的填补矩阵, 记作$X_{\mathrm{old}}^{\mathrm{imp}}$;

对于$s\in M$

对于$x_{\mathrm{obs}}^{(s)}$与$y_{\mathrm{obs}}^{(s)}$利用随机森林的方法分类;

根据分类结果, 利用$x_{\mathrm{mis}}^{(s)}$预测$y_{\mathrm{mis}}^{(s)}$;

利用得到的预测值$y_{\mathrm{mis}}^{(s)}$更新填补矩阵, 记作$X_{\mathrm{new}}^{\mathrm{imp}}$;

对s中其余缺失变量继续填补;

直到满足停止准则γ;

(iv) 得到最终填补矩阵, 记作X^{imp}.

上述的停止准则γ为: 如果新的填补矩阵与之前的填补矩阵的差别增加那么循环停止. 其中连续变量的差别为

$$\Delta_N = \Big[\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}^{\mathrm{new}} - x_{ij}^{\mathrm{old}})^2\Big]\Big/\Big[\sum_{j\in N}\sum_{i=1}^{n}(x_{ij}^{\mathrm{old}})^2\Big];$$

分类变量的差别为

$$\Delta_F = \Big[\sum_{j=1}^{p}\sum_{i=1}^{n} I_{\{x_{ij}^{\mathrm{new}} \neq x_{ij}^{\mathrm{old}}\}}\Big]\Big/(*NA),$$

这里$*NA$是分类变量里缺失数据的数量.

3.2　成分数据缺失值的随机森林填补法

随机森林填补方法优点在于可以处理多类型变量的数据, 并在高维情况下表现良好. 但是经过多次实验发现, 在缺失率较低的情况下, 直接在成分数据上使用随机森林法进行缺失值填补的误差率大于基于Aitchison距离的k近邻方法(见4.1节图2), 这种现象在欧氏空间中并不存在[13]. 为了解决这个问题, 本文提出了一种解决方法, 即先对成分数据进行ilr变换到欧式空间, 然后用3.1节的方法对缺失值进行填补, 再将填补后的数据经过ilr逆变换到单形空间上并做修正使其满足成分数据的性质. 具体步骤如下:

令

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^{\mathsf{T}} \\ \boldsymbol{x}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathsf{T}} \end{pmatrix} = \big(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(D)}\big) = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1D} \\ x_{21} & x_{22} & \ldots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nD} \end{pmatrix} \tag{5}$$

为含有n个观测值, D个部分的成分数据观测矩阵. $\boldsymbol{m} \subseteq \{1, 2, \ldots, n\}$记为含有缺失部分的行指示集, $\boldsymbol{g} \subseteq \{1, 2, \ldots, D\}$为含有缺失部分的列指示集.

(i) 选用k近邻或直接利用随机森林法得到的填补矩阵作为初始填补值, 且初始矩阵记作X_0.

(ii) 令$i = 1$.

(iii) 对每个初始填补值x_{ij} ($i \in \boldsymbol{m}$, $j \in \boldsymbol{g}$)而言, 将$\boldsymbol{x}_1$与$\boldsymbol{x}_i$交换; $\boldsymbol{x}^{(1)}$与$\boldsymbol{x}^{(j)}$交换, 即将每个缺失值调整到数据矩阵的第一个位置, 方便进行填补. 即: $X' = I_i X_0 I_j$, I_i与I_j分别是将单位矩阵的第i行与第1行交换, 第j列与第1列交换后的矩阵. 由于数据只进行了行列交换, 则数据的本身性质并未发生变化. 此时矩阵记作X'.

(iv) 利用等距对数比变换(ilr)将行列交换后的数据矩阵X'由单形空间变换到欧式空间, 记作Z. 即:

$$\begin{aligned}\mathrm{ilr}(X') &= \mathrm{ilr}\begin{pmatrix} x_{ij} & x_{i2} & \dots & x_{i1} & \dots & x_{iD} \\ x_{2j} & x_{22} & \dots & x_{21} & \dots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{1j} & x_{12} & \dots & x_{11} & \dots & x_{1D} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{nj} & x_{n2} & \dots & x_{n1} & \dots & x_{nD} \end{pmatrix} \\ &= \mathrm{ilr}\begin{pmatrix} X_1' \\ B \end{pmatrix} = \begin{pmatrix} Z_{ij} & \dots & Z_{i(D-1)} \\ \multicolumn{3}{c}{\mathrm{ilr}(B)} \end{pmatrix} \doteq Z,\end{aligned}$$

其中, z_{ij}是经过ilr变换后的x_{ij}.

(v) 令z_{ij}为缺失, 利用3.1节随机森林法对z_{ij}进行填补, 得到填补值为z_{ij}^*. 由于此时只有一个缺失值, 则单个填补的效果更为显著, 此时数据矩阵记作Z^*, 且

$$Z^* = \begin{pmatrix} z_{ij}^* & z_{i2} & \dots & z_{i(D-1)} \\ \multicolumn{3}{c}{\mathrm{ilr}(B)} \end{pmatrix}.$$

(vi) 利用ilr逆变换公式将Z^*还原成成分数据, 由于z_{ij}的更新, 则第一行经过逆变换后也全部发生改变, 逆变换后的矩阵记作X'^*, 且

$$\mathrm{ilr}^{-1}(Z^*) = \begin{pmatrix} x_{ij}^* & x_{i2}^* & \dots & x_{iD}^* \\ \multicolumn{4}{c}{B} \end{pmatrix} = \begin{pmatrix} X_1'^* \\ B \end{pmatrix} = X'^*.$$

(vii) 由于原始成分数据每行定和不同, 而$X_1'^* = (x_{ij}^*, x_{i2}^*, \dots, x_{iD}^*)$是一组和为1的成分数据, 则需要对填补值$x_{ij}^*$进行修正, 修正因子为

$$f_{ij} = \frac{\mathrm{median}_{k\neq s} x_{ik}}{\mathrm{median}_{k\neq s} x_{ik}^*}, \tag{6}$$

则最终填补值为

$$\widehat{x}_{ij}^* = f_{ij} x_{ij}^* = \frac{\mathrm{median}_{k\neq s} x_{ik}}{\mathrm{median}_{k\neq s} x_{ik}^* x_{ij}^*}. \tag{7}$$

(viii) 对于$j = 2, 3, \dots, D$重复步骤(iii)–(vii).

(ix) 对于$i = 2, 3, \dots, n$重复步骤(ii)–(viii)依次得到每个缺失位置的对应填补值.

由步骤(viii)和(ix)可以看出, 这是一个迭代插补的过程. 这样既可以充分利用随机森林填补法在欧氏空间上填补的优势, 又能对每一个缺失值逐个插补, 依次更新数据矩阵, 提高随机森林法对成分数据缺失值填补的准确率.

3.3　评价标准

由于成分数据的特殊几何性质, 只有Aitchison距离可以明确表示出两个成分数据之间的差距, 因此我们选择成分平均误差(compositional mean error, cme)作为评价标准, 即:

$$\mathrm{cme} = \frac{1}{n_{\boldsymbol{m}}} \sum_{i \in \boldsymbol{m}} d_A(\boldsymbol{x}_i, \widehat{\boldsymbol{x}}_i^*), \tag{8}$$

其中$\boldsymbol{x}_i$是原始成分, $\widehat{\boldsymbol{x}}_i^*$是经过填补后的成分, $n_{\boldsymbol{m}}$是观测值的缺失部分数. cme用来计算原始数据和填补后数据的平均Aitchison距离.

§4.　模拟和实例

本节利用提出的新方法针对高维数低样本量和常见的低维数高样本量的成分数据进行缺失值填补以验证所提方法的有效性.

4.1　模拟

为了模拟在单形空间上的成分数据, 根据定义1, 利用等距对数比逆变换, 运用R语言编程将服从正态分布的欧氏数据逆变换出不同维数、相关性的成分数据, 见表1, 其中$\mathbf{1}_n = (1, 1, \ldots, 1)^\intercal$, r为相关系数, 本文分别取$r = 0.1, 0.5, 0.9$表示不同的相关性. 调整数据的缺失率为5%–30%且缺失类型为完全随机缺失(MCAD), 利用随机森林法(RF)、变换后的随机森林法(IRF) (3.2节插补过程中不做中位数修正)、变换后并做中位数修正的随机森林法(MIRF)、基于Aitchison距离的k近邻法(KNN)分别对缺失值进行填补并并计算对应的成分平均误差cme.

表1　两组模拟数据的参数设定

数据	分布	维数$(D-1)$	观测值(n)	μ	Σ
X_1	$X_1 \sim \mathrm{N}_s^{51}(\mu_1, \Sigma_1)$	50	10	$\mu_1 = (0, 0, \ldots, 0)$	$\Sigma_1 = r\mathbf{1}_{10}\mathbf{1}_{10}^\intercal + (1-r)I_{10}$
X_2	$X_2 \sim \mathrm{N}_s^{11}(\mu_2, \Sigma_2)$	10	50	$\mu_2 = (0, 0, \ldots, 0)$	$\Sigma_2 = r\mathbf{1}_{50}\mathbf{1}_{50}^\intercal + (1-r)I_{50}$

图1中, 基于Aitchison距离的k近邻法由于设置成分数据的维数过高且观测值较少, 随着缺失率的增大导致整列缺失, 不能找到与目标样本最接近的其他样本, 程序无法运行, 因此只是比较了其余三种填补方法. 由图1和图2可见, 随着成分数据相关性的减小和缺失率的增大, 四种填补方法的成分平均误差cme都在增大, 这是必然的. 经过两种不同类型数据的模拟可见无论维数、相关性和缺失率的变化, 成分数据经过ilr变换并做中位数修正的随机森林法(MIRF)的成分平均误差cme明显小于其他三种方法, 这验证了新方法广泛的适用性和填补效果的准确性.

下面我们通过实例分析进一步说明经过ilr 变换后并做中位数修正的随机森林法(MIRF)的有效性.

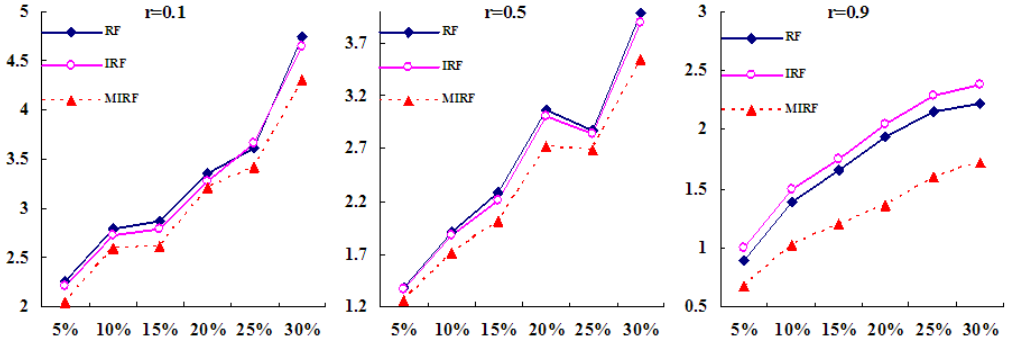

图1 模拟高维成分数据X_1在不同相关性和缺失率下三种方法填补效果的比较
(横坐标为缺失率, 纵坐标为成分平均误差cme)

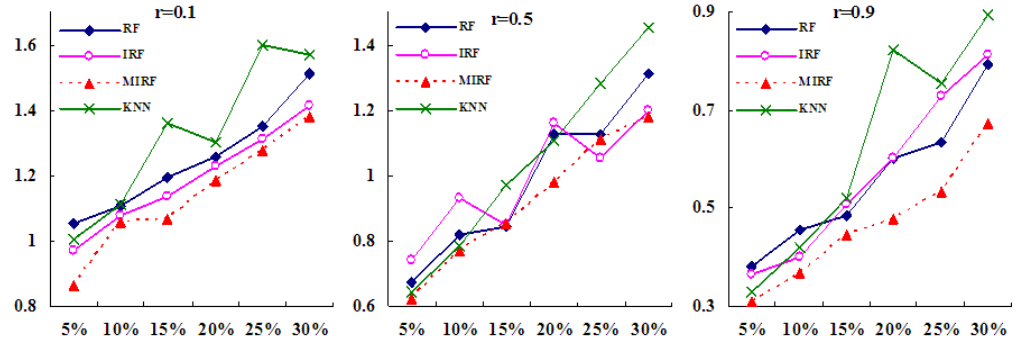

图2 模拟普通成分数据X_2在不同相关性和缺失率下四种方法填补效果的比较
(横坐标为缺失率, 纵坐标为成分平均误差cme)

4.2 实例

实例1: 这里选用的数据集来自文献[5; P. 354]的Hongite岩石矿物成分数据, 这个数据集包含25个样本, 每个样本包含5种矿物成分变量, 分别标记为A、B、C、D、E. 这个数据集是完整的, 利用R语言软件随机选取10个值并设置其缺失, 再使用随机森林法(RF)、变换后的随机森林法(IRF)、变换后并做中位数修正的随机森林法(MIRF)和基于Aitchison距离的k近邻法(KNN)进行填补, 最终将填补得到的成分平均误差(cme)和填补值与原始真实值作比较. 结果见表2.

表2 Hongite岩石矿物成分数据下四种方法填补值的比较

Method	cme	1	2	3	4	5	6	7	8	9	10
真实值	0	34.6	13.2	16.8	25	11.4	10.1	10.9	24.9	1	9.4
RF	0.1802	35.06	10.79	15.54	**25.54**	13.56	10.93	**10.97**	22.54	2.75	8.19
IRF	0.2049	**34.96**	10.51	12.43	27.06	14.33	10.89	11.03	23.23	2.70	**8.69**
MIRF	**0.1740**	30.13	9.89	**17.91**	21.42	**13.43**	**10.73**	10.66	**23.32**	**2.01**	8.32
KNN	0.3953	27.74	**11.33**	29.47	22.02	15.53	8.71	7.48	12.19	3.22	6.73

由表2可知, 对于低维数的实际数据集而言, 数据经过ilr变换后并做中位数修正的随机森林法(MIRF)的填补效果同样明显优于原始的随机森林法, 且基于Aitchison距离的k近邻法(KNN)效果最不理想.

实例2: 此时选用的数据集为来自中国统计年鉴[14]的农牧渔业生产情况数据, 这个数据集包含2007年–2013年的农牧渔业生产情况共6个样本, 每个样本包含27种农产品单位面积产量. 这个数据集是完整的, 由于维数的增大, 过高的缺失率会导致数据失真, 本实例利用R语言软件随机选取2%–10%的值并设置其缺失, 再使用不同的方法进行填补, 将得到的成分平均误差(cme)做比较. 结果见表3.

表3　农牧渔业生产情况数据下三种填补方法成分平均误差(cme)比较

方法	2%	4%	6%	8%	10%
RF	1.888	1.707	3.065	3.034	3.309
IRF	1.320	1.652	3.127	3.028	3.204
MIRF	**0.692**	**1.146**	**1.955**	**2.467**	**2.225**

由表3可知, 数据经过ilr变换后并做中位数修正的随机森林法(MIRF)进行缺失值的填补在高维实际数据中也明显适用, 且准确率最高.

§5. 结　　论

大多数统计方法都是建立在完整数据集上, 因此缺失值的处理是有必要的, 对于高维低样本量的数据而言许多现有的填补方法并不适用. 此外, 由于成分数据的特殊几何性质“正定性”和“定和性”, 需要利用非参数方法或对数据进行ilr变换到欧氏空间. 为了估计成分数据的缺失值, 本文提出了一种基于随机森林的成分数据缺失值迭代填补法, 并在模拟和实例中选取不同类型的数据将新方法与原始的随机森林法和基于Aitchison距离的k近邻填补法进行比较, 证明了新方法的有效性. 本文选取的中位数修正方式可以改变, 根据数据类型和维数的不同在未来的研究中也许可以找到更有效的修正方法对成分数据的缺失值进行填补.

参 考 文 献

[1] García-Laencina P J, Sancho-Gómez J L, Figueiras-Vidal A R, et al. *K* nearest neighbours with mutual information for simultaneous classification and missing data imputation [J]. *Neurocomputing*, 2009, **72(7-9)**: 1483–1493.

[2] Wang H, Wang S H. Discovering patterns of missing data in survey databases: an application of rough sets [J]. *Expert Syst. Appl.*, 2009, **36(3)**: 6256–6260.

[3] Little R J A, Rubin D B. *Statistical Analysis with Missing Data* [M]. 2nd ed. New Jersey: Wiley, 2002.

[4] Ferrers N M. *An Elementary Treatise on Trilinear Co-Ordinates* [M]. London: Macmillan, 1866.

[5] Aitchison J. *The Statistical Analysis of Compositional Data* [M]. New York: Chapman and Hall, 1986.

[6] Aitchison J, Barceló-Vidal C, Martín-Fernández J A, et al. Logratio analysis and compositional distance [J]. *Math. Geol.*, 2000, **32(3)**: 271–275.

[7] Egozcue J J, Pawlowsky-Glahn V, Mateu-Figueras G, et al. Isometric logratio transformations for compositional data analysis [J]. *Math. Geol.*, 2003, **35(3)**: 279-300.

[8] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm [J]. *J. Roy. Statist. Soc. Ser. B*, 1977, **39(1)**: 1–38.

[9] Hron K, Templ M, Filzmoser P. Imputation of missing values for compositional data using classical and robust methods [J]. *Comput. Statist. Data Anal.*, 2010, **54(12)**: 3095–3107.

[10] Han J W, Kamber M. *Data Mining: Concepts and Techniques* [M]. 2nd ed. San Diego, USA: Academic Press, 2006.

[11] Martín-Fernández J A, Barceló-Vidal C, Pawlowsky-Glahn V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation [J]. *Math. Geol.*, 2003, **35(3)**: 253–278.

[12] Breiman L. Random forests [J]. *Mach. Learn.*, 2001, **45(1)**: 5–32.

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

[14] 中华人民共和国国家统计局. 中国统计年鉴 [M]. 北京: 中国统计出版社, 2014.

Imputation of Missing Values for Compositional Data Based on Random Forest

ZHANG XiaoQin CHENG YuYing

(*School of Mathematics Sciences, Shanxi University, Taiyuan, 030006, China*)

Abstract: Dealing with the missing values is an important object in the field of data mining. Besides, the properties of compositional data lead to that traditional imputation methods may get undesirable result if they are directly used in this type of data. As a result, the management of missing values in compositional data is of great significant. To solve this problem, this paper uses the relationship between compositional data and Euclidean data, and proposes a new method based on Random Forest for missing values in compositional data. This method has been implemented and evaluated using both simulated and real-world databases, then the experimental results reveal that the new imputation method can be widely used in various types of data sets and has good performance than other methods.

Keywords: imputation of missing values; compositional data; random forest

2010 Mathematics Subject Classification: 62G05; 62P05