DOI: 10.11992/tis.201906015
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190902.1033.002.html

构造性覆盖下不完整数据修正填充方法

严远亭，吴亚亚，赵姝，张燕平
（安徽大学 计算机科学与技术学院，安徽 合肥 230601）

摘　要：不完整数据处理是数据挖掘、机器学习等领域中的重要问题，缺失值填充是处理不完整数据的主流方法。当前已有的缺失值填充方法大多运用统计学和机器学习领域的相关技术来分析原始数据中的剩余信息，从而得到较为合理的值来替代缺失部分。缺失值填充大致可以分为单一填充和多重填充，这些填充方法在不同的场景下有着各自的优势。但是，很少有方法能进一步考虑样本空间分布中的邻域信息，并以此对缺失值的填充结果进行修正。鉴于此，本文提出了一种可广泛应用于诸多现有填充方法的框架用以提升现有方法的填充效果，该框架由预填充、空间邻域信息挖掘和修正填充三部分构成。本文对 7 种填充方法在 8 个 UCI 数据集上进行了实验，实验结果验证了本文所提框架的有效性和鲁棒性。

关键词：不完整数据；缺失值填充；邻域信息；数据挖掘；机器学习；填充方法；单一填充；多重填充
中图分类号：TP18　　**文献标志码**：A　　**文章编号**：1673−4785(2019)06−1225−08

中文引用格式：严远亭，吴亚亚，赵姝，等. 构造性覆盖下不完整数据修正填充方法 [J]. 智能系统学报，2019，14(6)：1225–1232.
英文引用格式：YAN Yuanting, WU Yaya, ZHAO Shu, et al. Improving missing data recovery with a constructive covering algorithm[J]. CAAI transactions on intelligent systems, 2019, 14(6): 1225–1232.

Improving missing data recovery with a constructive covering algorithm

YAN Yuanting, WU Yaya, ZHAO Shu, ZHANG Yanping
(School of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract: Incomplete data processing is one of the most active avenues in the fields of data mining, machine learning, etc. Missing value imputation is the mainstream method used to deal with incomplete data. At present, most existing missing value imputation methods utilize relevant techniques in the field of statistics and machine learning to analyze surplus information from original data to replace the missing attributes with plausible values. Missing value imputation can be roughly divided into single imputation and multiple imputation, which have their own advantages in different scenarios. However, there are few methods that can further consider neighborhood information in the spatial distribution of samples and modify the filling results of missing values. In view of this, this paper proposes a new framework that can be widely used in many existing imputation methods to enhance the imputation effect of existing methods. It is composed of three modules, called pre-filling, spatial neighborhood information mining, and modification of the results of pre-filling separately. In this paper, seven existing imputation methods were evaluated on eight UCI datasets. Experimental results verified the validity and robustness of the framework proposed in this paper.

Keywords: incomplete data; missing value imputation; neighborhood information; data-mining; machine learning; imputation method; single imputation; multiple imputation

收稿日期：2019−06−06.　**网络出版日期**：2019−09−02.
基金项目：国家自然科学基金项目 (61806002, 61872002, 61673020, 61876001, 61602003)；安徽省自然科学基金项目 (1708085QF143, 1808085MF197)；安徽大学博士科研启动基金项目 (J01003253).
通信作者：张燕平. E-mail: zhangyp2@gmail.com.

机器学习、数据挖掘等技术在诸如生物特征识别、文本分类和医学诊断等领域得到了广泛应用[1-6]。近年来，随着传感器技术、信息技术等科学技术的迅猛发展，数据获取的途径日益丰富，

这给机器学习等技术带来了极大的发展机遇。然而在实践中，通常会因为存储设备损坏、数据采集设备能力有限等多种因素导致数据出现缺失的情况，我们称其为不完整数据问题。此类问题普遍存在于众多领域中，如：微阵列数据[7-8]、移动电话数据[9]、可视化数据[10]、工业数据[11]、软件项目数据[12]等。然而，传统机器学习的方法往往都是针对完整数据而设计的，因此缺失数据给这些方法带来了极大挑战。

目前已有不少学者针对不完整数据提出了一些解决策略，大致可以分为三类，其一为替代法，即用同一数据集内其他样本的完整部分替代缺失值，有时甚至会将众多缺失属性补以统一的固定值。这种策略虽然简单，但众多研究表明，绝大多数原始数据集的样本属性间都不是相互独立的，因此单一的替换策略直接忽略了属性间的关系，并不可取；其二为删除法，例如在许多统计软件如：SPSS、SAS 中，默认采用 Listwise deletion(LD) 策略处理缺失值，直接删除带有缺失项的样本。然而，这种策略是以减少原始数据为代价换取数据完整，在信息获取代价较大时会造成严重的资源浪费和重要信息的损失。因此，解决不完整数据问题的第 3 种策略，即在多数现有的机器学习算法被应用到实际问题之前，将缺失数据填充完整的策略更为主流一些。

当前的缺失值填充方法大多运用统计学和机器学习领域的相关技术，对不完整数据的剩余部分进行建模和分析，从而产生较为合适的值用以填充。最常用的统计学填充法是均值填充[13]，它简单快速，但是无法较好地拟合原始数据，因此通常适用于快速填充或者只有极少数属性缺失的情况；同样基于统计学的回归填充通常基于数据的完整部分来建立回归模型，对于包含缺失值的样本，将已知属性值代入方程来估计未知属性值。除此以外，在过去的十年里，许多机器学习填充方法也被相继提出，在机器学习填充法中，缺失的属性通常被视为一个训练模型的目标输出，剩余其他完整属性是用于训练和测试的输入特性，算法通常根据数据集的完备部分使用一些诸如 KNN、决策树 (DT)、多层感知器 (MLP)、自组织映射 (SOM) 等机器学习方法来训练相关模型，在模型中对不完整属性进行估计。

当前最常用的机器学习填充方法有 K 最近邻填充 (KNNI)[14]、聚类填充 (CKNNI[15]) 等。其中，K 最近邻填充的一个最大的特点在于 KNN 是一种懒惰式的学习方法，在应用到缺失值填充问题时不需要建立明确的模型 (如决策树模型或诸多其他繁琐的填充规则)，无论使用何种方法对数据进行分析，总能找到距缺失样本最近的若干个样本用于填充。因此，基于该方法的很多改进方法也被相继提出，如 WKNNI[16] 等，它们通常使用某种度量指标 (如皮尔逊相关系数、距离函数等) 来衡量样本间的相似度大小或采用特征提取等手段来对重要属性进行加权后再进行填充，取得了较好的填充效果；聚类填充方法如：CKNNI[15] 根据样本间的相似度大小将所有样本聚类得到若干个簇，簇内样本间的相似度较高，再针对缺失样本，利用其所在簇内的其他样本对缺失部分进行填充。虽然这类方法也在一定程度上考虑了样本在空间分布中的信息，但由于聚类是无监督的，因此无法知悉经过聚类操作后会得到怎样的聚类结果，同时聚类算法初始中心个数也很难确定。此外，还有一些其他的机器学习算法被应用于缺失值填充领域，如结合期望最大化方法 (EM) 用于最大似然估计的方法、基于支持向量机的填充方法等，但是这些方法在时间复杂度方面都非常庞大，收敛速度极慢，对于某些数据集甚至达到了指数级别。

上述所提方法大多为单一填充法 (如均值填充、KNN 填充等)，往往会降低估计量的方差。针对这一缺陷，Rubin 在 1987 年提出了多重填充的思想。单一填充往往针对每个缺失值产生一个可能的值用以填充，而多重填充 (如 MICE[17]、MILC[18-19]) 是指在对填充值的预测分布中，通过一组 (m>1) 合理的值来替代所有缺失值的过程[5, 20]。数据经过多重填充处理后，会得到 m 个完整的数据集，每一个数据集都可以运用分析完整数据的方法对其分析，然后再融合这些不同数据集的分析结果，给出综合估计，显著缩小了由单一填充所导致的偏差，可获得更好的填充效果。

尽管现有的这些填充方法都具有各自的优势，某些方法能较好地实现对缺失数据的恢复。但是研究表明，目前还没有哪一种填充方法可以在任意给定的数据集和所有场景下都取得最佳的填充效果[21]。此外，现有的绝大多数方法仍缺乏对样本空间分布信息的考虑，忽略了空间邻域信息对数据恢复的影响。因此，本文提出了一种新的框架，可用于诸多现有的填充方法以进一步提升填充效果。框架由 3 部分组成，分别为预填充、空间邻域信息挖掘和修正填充。首先利用传统的填充方法对数据集进行预填充，得到完整数据；然后构造性的对预填充后的数据集构造覆

盖, 挖掘样本的空间邻域信息; 最后利用邻域内样本的有效信息对预填充的结果进行修正, 从而得到最终的完整数据集。

1 相关工作

1.1 缺失机制和缺失模式

根据 Rubin 的总结, 共有 3 种类型的缺失机制会导致一个不完整数据集的产生, 分别为完全随机缺失、随机缺失和非随机缺失。

完全随机缺失: 样本的某一属性出现缺失的概率和其他样本以及该属性本身的值无关。即当某属性值发生缺失的可能性与其他样本无关且与该样本属性自身也无关时称作完全随机缺失。

随机缺失: 当某样本属性缺失的可能性与模型中某些观测样本有关而与该样本自身无关时称其为随机缺失。

非随机缺失: 当样本属性中存在缺失值的可能性仅与其自身相关时则称为非随机缺失。

除此之外, 还有两种缺失模式, 分别为单调缺失和非单调缺失。前者指的是针对同一个记录或者变量的缺失, 后者指的是针对任何记录、任何变量的缺失, 本文的实验是以非单调缺失模式为前提进行的。

1.2 构造性覆盖算法

张铃等[22] 提出的基于覆盖的构造性机器学习方法主要是以 M-P 模型的几何表示为理论基础, 针对样本自身的特点来构造神经网络。构造性覆盖算法可以看作一个 3 层网络分类器。

输入层: 共 n 个神经元, 每个神经元对应样本的一维, 即样本的特征属性, $x_i=(x_i^1,x_i^2,\cdots,x_i^n)$, 该层神经元只负责接收外部信息, 自身无信息处理能力。

隐藏层: 共 s 个神经元。初始时, 隐层神经元为 0 个, 每求得一个球形覆盖, 增加一个神经元, 直到将所有的样本都被覆盖, 从而求得一组覆盖: $C=\left\{C_1^1,C_1^2,\cdots,C_1^{n_1},C_2^1,C_2^2,\cdots C_2^{n_2},\cdots,C_1^m,\cdots C_m^{n_m}\right\}$, 其中 C_i^j 表示第 i 类样本的第 j 个覆盖, 是隐层中的一个神经元, 隐层共有 $s=\sum n_i$ 个覆盖, 第 i 类有 n_i 个覆盖, $i=1,2,\cdots m$。神经元的权值是覆盖的中心, 阈值为覆盖的半径。

输出层: 共 m 个神经元, 第 t 个神经元的输入为同类的一组覆盖, 输出为该覆盖的类别。$O_t=(O_1=0,O_2=0,\cdots,O_t=1,\cdots,O_m=0)$ 表示第 t 类样本的输出。该层神经元向外部输出处理信息。构造性覆盖算法属于有监督学习。

构造性覆盖算法针对样本自身的特点, 根据学习样本的特征用超平面切割球面形成“球形领域”作为神经元来构造神经网络, 基于超球面上的样本来构造每个类别的覆盖。该算法可以处理海量样本, 适用于多分类问题且分类能力强、运算速度快[23]。

2 基于空间邻域信息的修正填充方法

2.1 预填充

给定不完整数据集 $\mathrm{DS}=(x_i,y_i|i=1,2,\cdots,m)$, 令 $F=\{F^{(1)},F^{(2)},\cdots,F^{(n)}\}$。其中, m 表示的是样本个数; F 表示的是输入空间的特征集合; $x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})$; $x_i^{(j)}$ 表示的是第 i 个样本的第 j 维属性, 若第 i 个样本的第 j 维缺失, 则记 $x_i^{(j)}=NaN$。定义所有缺失值的集合:

$$\mathrm{IND}=\{(i,j)|x_i^{(j)}=NaN\} \tag{1}$$

利用经典的填充算法对 DS 进行填充, 在本文中称为预填充。预填充后得到完整数据集 DS_c。分别选用以下 7 种填充方法实现预填充:

1) 均值填充 (MEI)[13]

$$x_i^{(j)}=\sum_{k\in I(\mathrm{comp})}\frac{x_k^{(j)}}{n_{|I(\mathrm{comp})|}} \tag{2}$$

式中: $I(\mathrm{comp})$ 表示的是所有第 j 维属性不缺失的属性索引集合; 而 $n_{I(\mathrm{comp})}$ 表示的是所有第 j 维属性不缺失的样本总数。

2) 中值填充 (MEDI)[13]

令 $\Omega=\left\{x_k^{(j)}|k\in I(\mathrm{comp})\right\}$; Ω^* 表示集合 Ω 中所有元素的升序存放; $|\Omega|=|\Omega^*|=n$; 则:

$$x_i^j=\begin{cases}\Omega^*_{[n+1/2]}, & n\ \bmod 2\neq 0\\ (\Omega^*_{[n/2]}+\Omega^*_{[n+1/2]})/2, & n\ \bmod 2=0\end{cases} \tag{3}$$

3) KNNI[14]

$$d_i^t=\mathrm{dist}(x_i,x_t)=\sqrt{\sum_{p=1}^{n}(x_i^p-x_t^p)^2} \tag{4}$$

$$x_i^j=\left(\sum_{\forall x_t\in N_i}x_t^j\right)/|N_i|,\quad x_t^j\neq \mathrm{NaN} \tag{5}$$

式中: d_i^t 表示第 i 个样本和第 j 个样本之间的欧式距离; N_i 表示距离第 i 个样本最近的 k 个第 j 维属性不缺失的样本集合; $|N_i|=k$ 表示的是 N_i 集合中的个数。

4) WKNNI[16]

$$x_i^j=\sum_{\forall x_t\in N_i}x_t^j\omega_t,\quad \omega_t=\frac{1/d_i^t}{\sum_{t=1}^{k}1/d_i^t} \tag{6}$$

式中: N_i 表示距离第 i 个样本最近的 k 个第 j 维属性不缺失的样本集合; ω_t 表示第 t 个样本的权值; d_i^t 表示第 i 个样本和第 j 个样本之间的欧式距离;

k 表示最近邻的个数。

5) Soft-impute(SoftI)[24]: 通过对 SVD 分解的迭代软阈值处理来填充不完整数据。

6) Matrix-Factorization-impute(MFI)[25]: 将不完整数据用矩阵形式表示并直接将其分解为低秩的 $\boldsymbol{U}$ 和 $\boldsymbol{V}$, 然后对 $\boldsymbol{U}$ 中的元素采用 L1 稀疏惩罚, 对 $\boldsymbol{V}$ 中的元素采用 L2 稀疏惩罚, 通过梯度下降法求解。

7) MICE[20]: 利用链式方程实现多重填充。

2.2　挖掘空间邻域信息

在对不完整数据进行填充的过程中,最关键的问题在于如何通过对数据集中样本的剩余完整信息进行分析。在这一节中, 我们利用一种有监督的空间邻域信息挖掘方法, 挖掘与缺失样本具有更高相似性的某邻域内样本的有效信息。

1) 通过式 (2) 变换将 DS_c 中的样本点投射到 S^{n+1} 球面上并使得投影后的样本向量等长。

$$T:\mathrm{DS}_c\to S^{n+1}, T(x)=\left(x,\sqrt{R^2-|x|^2}\right) \tag{7}$$

式中$R\geqslant\max\{|x|,x\in DS_c\}$

2) 随机选取一个未被标记的样本 x_k 作为覆盖中心并计算覆盖半径 R。

$$<x_k,x_i>x_k^{(1)}x_i^{(1)}+\cdots+x_k^{(n+1)}x_i^{(n+1)},\quad i\in\{1,2,\cdots,m\} \tag{8}$$

$$d_1=\max_{y_k\neq y_i}\{\langle x_k,x_i\rangle\},\quad i\in\{1,2,\cdots,m\} \tag{9}$$

$$d_2=\min_{y_k=y_i}\{\langle x_k,x_i\rangle|\langle x_k,x_i\rangle>d_1\},\quad i\in\{1,2,\cdots,m\} \tag{10}$$

$$R=(d_1+d_2)/2 \tag{11}$$

式中: $<x_k,x_i>$ 表示样本 x_k 与 x_i 之间的内积; d_1 表示异类样本间的最小距离, 等价于最大内积; d_2 表示同类样本间的最大距离, 等价于最小内积; R 表示覆盖半径。

3) 构造一个以 x_k 为球形领域的中心, R 为半径的球形领域 C_v, 其中 C_v^k 表示第 v 类样本的第 k 个覆盖, 并将该领域内的所有样本都标记成“已学习”。若全部已被标记则会得到一组覆盖集合 $C=\{C_1^1,C_1^2,\cdots,C_1^{k_1},C_2^1,C_2^2,\cdots,C_2^{k_2},\cdots,C_i^1,\cdots\}$, 否则返回 2)。

经过 1)~3) 得到的覆盖集合 C 能够很好地刻画样本空间的空间邻域信息。

2.3　利用空间邻域信息修正预填充结果

为方便起见, 本文以二分类问题为例, 描述如何利用空间邻域信息进行缺失值的填充。

令 C_1 和 C_2 为经过 2.2 节所得到的两类样本的覆盖集合为

$$C_1=\left\{C_1^1,C_1^2,\cdots,C_1^{k_1}\right\};C_2=\left\{C_2^1,C_2^2,\cdots,C_2^{k_2}\right\} \tag{12}$$

式中: k_1 表示第一类样本的覆盖的个数, k_2 表示第二类样本的覆盖的个数。

对于 $\forall x_i\in C_v^k\wedge(i,j)\in\mathrm{IND}$, 使用式 (13) 对 $x_i^{(j)}$ 进行修正填充。

$$x_i^{(j)}=\begin{cases}\left(\sum\limits_{t=1}^{|\psi|}(x_s^{(j)})_t\right)/|\psi|, & |\psi|\neq 0\\ x_p^{(j)}, & |\psi|=0\end{cases} \tag{13}$$

$$\psi=\left\{x_s^{(j)}|x_s\in\left\{C_v^k-x_i\right\}\wedge(s,j)\notin\mathrm{IND}\right\} \tag{14}$$

式中: ψ 表示在 C_v^k 覆盖集合内, 除样本 x_i 以外的所有第 j 维不缺失的属性值集合。$|\psi|$ 为满足条件 (14) 的所有属性的数量。

在极少数情况下会出现覆盖内除 x_i 以外, 其他所有样本对应的第 j 维属性在预填充步骤前均是缺失的, 即出现式 (13) 中的 $|\psi|$ 值为零的情况, 这表明这些属性在预填充阶段都已被填充, 算法采用预填充的结果 $x_p^{(j)}$ 替代 $x_i^{(j)}$, 其中 $(p,j)\in\mathrm{IND}$。

算法针对所有的缺失属性依次进行判断和修正填充, 最终得到修正填充后的完整数据集 DS_{fc}。修正填充方法的流程图如图 1 所示, 其中, MR 表示缺失率。

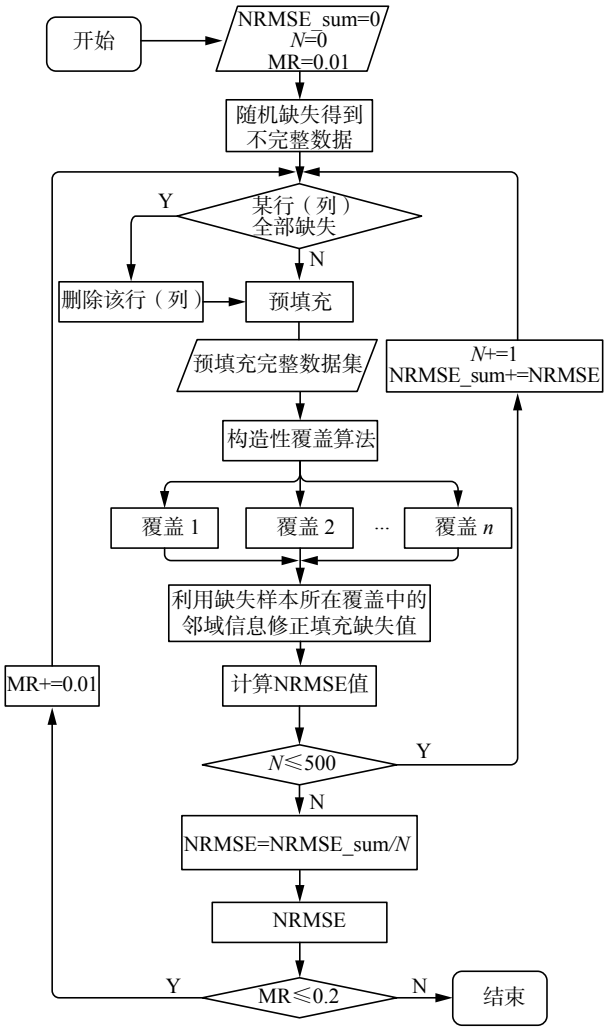


图 1　算法流程图

Fig. 1　The flow chart of the proposed method

3 实验与分析

3.1 实验设计

本小节中, 本文提出的框架分别与 2.1 节提出的 7 种当前已有的缺失值填充方法对比。由于本文的方法是对已有填充方法得到的填充结果进一步的修正。因此, 这 7 种经典的方法也是本文预填充阶段所选取的填充方法。另外, 本文使用一种常见的度量缺失值填充效果的指标 (NRMSE) 来量化填充效果, 进一步验证方法的有效性。

实验首先在 UCI 上的完整数据集上, 以随机缺失的方式得到不完整数据, 缺失率从小到大依次为 0.01、0.02、0.03、0.04、0.05、0.10、0.15 和 0.20; 然后用现有的填充方法与本文提出的框架结合后对得到的不完整数据进行处理; 最后根据修正填充前的原始数据集和修正填充后得到的最终完整集得出不同缺失率下的 NRMSE 值及其变化趋势。为了避免因单次填充而导致的误差对实验结果的影响, 实验在每种缺失率下均重复随机缺失多次, 最终得到的 NRMSE 值为多次重复缺失后得到的均值。

3.2 实验数据集

本文从 UCI 中选取了 8 个数据集进行实验的比较和分析, 表 1 给出了 8 个数据集的基本信息, 包括数据集名称、样本个数、属性个数以及类别个数, 其中 balance-s、BCC、dba 和 lym 分别是 balance-scale、Breast Cancer Coimbra、data_banknote_authentication 和 lymphography 数据集的简写。

表 1　数据集简介

Table 1　The introduction of the datasets

数据集	样本数	属性数	类别数
balance-s	625	4	3
BCC	116	10	2
dba	1 372	5	2
glass	214	10	7
haberman	306	3	2
lenses	24	4	3
lym	148	18	4
wine	178	13	3

3.3 不完整数据填充性能的评价标准

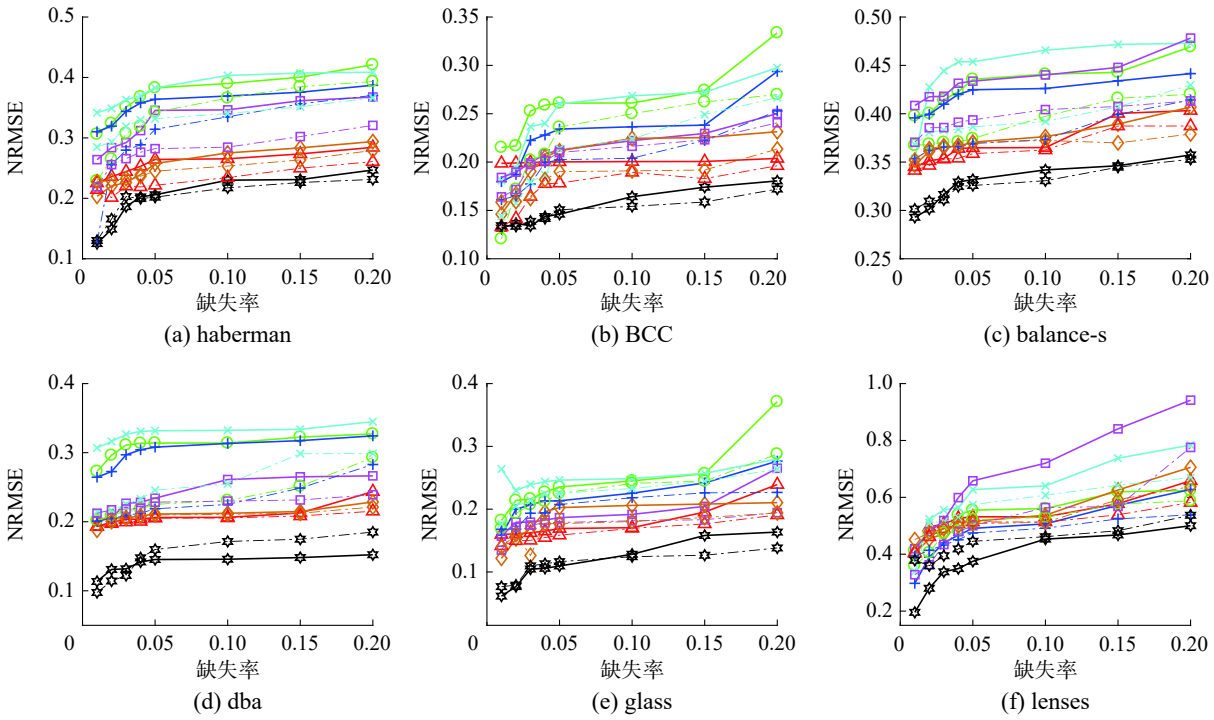
本文采用的评价不完整数据填充性能的指标是归一化均方根误差, 定义如下:

$$\mathrm{NRMSE}=\frac{\sqrt{\mathrm{mean}(x_{\mathrm{guess}}-x_{\mathrm{ori}})^2}}{\mathrm{std}(x_{\mathrm{ori}})} \tag{15}$$

式中: x_{guess} 和 x_{ori} 分别代表填充后的属性值以及在原始数据集中被填充前的属性值; $\mathrm{std}(x_{\mathrm{ori}})$ 表示的是被填充前的所有属性值的标准差。NRMSE 值越小, 意味着填充后的值与填充前的值差异越小, 即填充效果越好。

3.4 实验结果与分析

本小节中, 通过 NRMSE 度量指标对本文所提框架与当前已有的缺失值填充方法所产生的结果进行对比和分析, 研究在不同的缺失率下, 框架对于现有填充方法的提升效果呈现出何种趋势。图 2 分别展示了 7 组填充方法在不同数据集上以及不同缺失率下对应的 NRMSE 值的变化趋势。



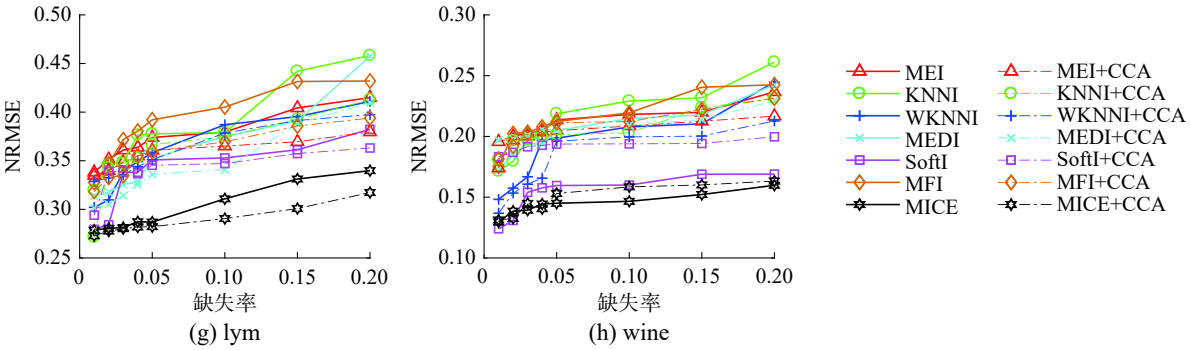


图 2　7 组填充方法在 8 个数据集上以及在不同缺失率下对应的 NRMSE 值

Fig. 2　The NRMSE values corresponding to 7 groups of imputation methods on 8 datasets and at different missing rates

除此之外，图 3 更为直观地展示了 7 种现有填充方法分别在与本文所提框架结合后填充效果的提升。其中，MR 表示的是缺失率。从实验结果中不难发现，由于数据的随机缺失导致缺失值的分布具有一定的随机性，同时因缺失场景的不同和缺失数据本身的特点共同决定了没有任何一种现有的填充方法可以在所有数据集上都能取得最好的填充效果。但是本文所提框架应用于多数现有填充方法和数据集上所呈现出来的填充效果均优于对比算法，如 haberman、BCC 等数据集。从总体上来看，针对一些特征 (维) 数较少而样本数较多的数据集而言，当缺失率较小时，对应的 NRMSE 值也较小，当缺失率达到最小时，NRMSE 值往往也达到最小，意味着填充效果最好，但是随着缺失率的不断增大，对应的 NRMSE 逐渐呈现出增长的趋势。这是因为缺失率越小，出现缺失的属性数就越少，而当样本数远大特征 (维) 数时，例如 dba 数据集的特征 (维) 数是 5，但却拥有 1 372 个样本，当缺失率为 1% 时，出现缺失的属性数是 13，而该数据本身具有 1 372 个样本，故可用于填充的剩余完整样本数还有很多，这对于数据的恢复十分有利。随着缺失率的增长，剩余完整样本逐渐变少，对应的 NRMSE 值也呈现出上升的趋势。当然，也会出现极少数的特殊情况，如：MEI 方法在 haberman 数据集上当缺失率为 2% 时，对应的 NRMSE 值略小于缺失率等于 1% 时对应的值，即随着缺失率的增大，NRMSE 值却在减小，类似的情况还有 MEDI 方法在 glass 数据集上以及 SoftI 方法在 balance-s 数据集上等。这种情况出现的原因是实验所得的不完整数据是通过随机缺失的方式得到的，且在挖掘空间邻域信息的过程中也采用了多次覆盖，因此具有一定的随机性，尽管重复随机缺失了很多次，最后取多次实验所得的 NRMSE 度量值的均值为最终结果，但也只能在一定程度上较为稳定地反映出 NRMSE 值的总体变化趋势，此外，针对一些数据样本较少的数据集而言，当缺失率较小时，对应缺失率下的缺失属性的数量也十分接近，在这种情况下，可能会出现缺失率略大反而填充效果略好的情况，但从经过多次重复实验的结果来看可知 NRMSE 指标的总体变化呈上升趋势。

从 BCC 和 balance-s 数据集上的实验结果中可以看出，尽管有少数对比算法在缺失率较小的时候呈现出的填充效果要优于本文的方法，但是，随着缺失率的不断增长，本文方法所展现出来的优势逐渐明显。如 MICE+CCA 在 balance-s 数据集上，当缺失率达到 5% 以及在 glass 数据集上达到 10% 以后，填充效果开始逐渐优于对比算法。此外，我们发现一些现有的单一填充方法在用本文框架修正填充后，填充效果更加逼近效果较好的多元填充方法，如：WKNNI+CCA 在 lenses 数据集上当缺失率达到 20% 时，几乎取得了等同于 MICE 的填充效果。这表明了通过和本文所提框架结合后可以更加有效地提升现有填充方法所产生的效果。

当然，也有方法与框架结合后的提升效果不佳，如：MICE+CCA 在 dba 和 wine 数据集上的填充效果较对比方法差。这是因为 MICE 本身就运用了多重填充的思想，已经在一定程度上规避了因单一填充而引入的误差，因此，在通常情况下都会取得十分不错的填充效果。但是，MICE+CCA 在 BCC、glass、haberman 和 lym 等数据集上随着缺失率的不断增大，对于 MICE 方法的提升效果逐渐明显。此外，同样是在 dba 和 wine 数据集上，本文所提出的用于修正填充其他现有方法的效果相对于对比算法的效果而言仍具有较大优势，如：KNNI+CCA、MFI+CCA 等。

从 lenses 数据集和 lym 数据集上的实验结果来看，一方面，除了 MICE 方法在 lenses 上的填充效果比 MICE+ CCA 好以外，其他几组对比算法的效果均没有本文提出的方法好，尤其是 SoftI+

CCA 在 lenses 数据集上以及 KNNI+CCA、MFI+CCA 等方法在 lym 数据集上的填充效果均显著优于对比算法; 另一方面, 尽管 MICE 在 lenses 上的效果更好一些, 但是与本文提出的 MICE+CCA 的效果相比较, 差距并不是很大, 而后者表现出的算法稳定性更高。

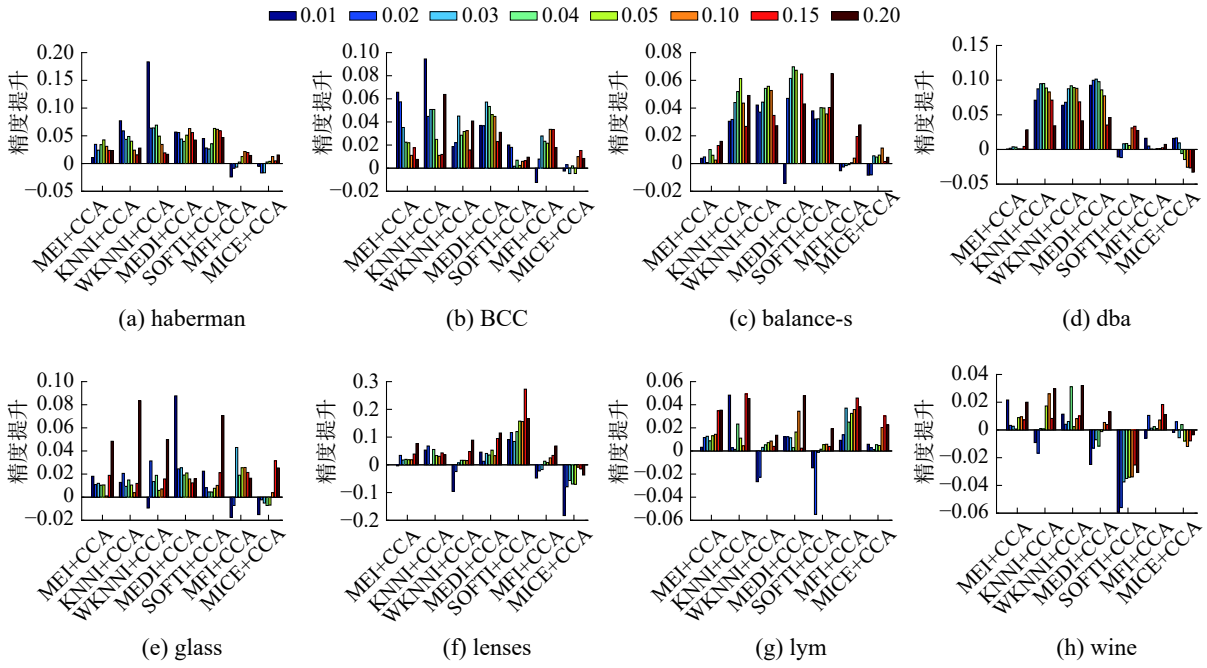


图 3　7 组现有填充方法经修正后的填充效果提升图

Fig. 3　The improvement of the imputation effect by using proposed method

4　结束语

针对当前已有的大多数填充方法忽视了样本空间分布信息对数据恢复的影响, 本文提出了一种可广泛应用于现有填充方法的框架, 旨在对利用现有方法得到的填充结果进行修正, 从而提升填充效果。该框架由预填充、空间邻域信息挖掘和修正填充 3 部分构成, 首先利用现有填充方法对样本进行预填充; 再通过引入一种空间邻域信息挖掘方法来找到与待填样本具有更高相似度的若干个空间邻域; 最后利用待填样本的空间邻域内的有效信息对现有填充方法产生的填充结果进行修正。

实验选取了 7 种经典的填充方法 (包括单一填充和多重填充), 在 8 个 UCI 数据集上进行了对比。结果表明, 本文提出的框架确实能够在大多数数据集上有效提升现有填充方法的填充效果。尽管在少数数据集上的提升效果不佳, 但是从实验所得的不同缺失率下的 NRMSE 度量值的变化趋势来看, 多数与框架结合后的填充方法通常呈现出较为平稳的填充趋势, 不会随着缺失率的不断增长而出现较大波动, 而且在某些数据集上呈现出一个重要规律, 即随着缺失率的不断增大, 框架对于现有填充方法的提升效果逐渐明显。除此之外, 本文所提框架还可以将一些效果较差的单一填充方法的填充效果提升至更好的多重填充方法所取得的效果。

参考文献:

[1] LARRAÑAGA P, CALVO B, SANTANA R, et al. Machine learning in bioinformatics[J]. Briefings in bioinformatics, 2006, 7(1): 86–112.

[2] HARPER P R. A review and comparison of classification algorithms for medical decision making[J]. Health policy, 2005, 71(3): 315–331.

[3] SEBASTIANI F. Machine learning in automated text categorization[J]. ACM computing surveys, 2002, 34(1): 1–47.

[4] KONG S G, HEO J, ABIDI B R, et al. Recent advances in visual and infrared face recognition—a review[J]. Computer vision and image understanding, 2005, 97(1): 103–135.

[5] FU Xiao, REN Yinzi, YANG Guiqiu, et al. A computational model for heart failure stratification[C]//Proceedings of 2011 IEEE Computing in Cardiology. Hangzhou, China, 2011: 385–388.

[6] FIALHO A S, KAYMAK U, ALMEIDA R J, et al. Probabilistic fuzzy prediction of mortality in intensive care units[C]//Proceedings of 2012 IEEE International Confer-

ence on Fuzzy Systems. Brisbane, Australia, 2012: 1–8.

[7] AITTOKALLIO T. Dealing with missing values in large-scale studies: microarray data imputation and beyond[J]. Briefings in bioinformatics, 2010, 11(2): 253–264.

[8] DE SOUTO M C P, JASKOWIAK P A, COSTA I G. Impact of missing data imputation methods on gene expression clustering and classification[J]. BMC bioinformatics, 2015, 16: 64.

[9] LIU Siyuan, CHEN Lei, NI L M. Anomaly detection from incomplete data[J]. ACM transactions on knowledge discovery from data, 2014, 9(2): 11.

[10] LIU Ji, MUSIALSKI P, WONKA P, et al. Tensor completion for estimating missing values in visual data[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1): 208–220.

[11] LAKSHMINARAYAN K, HARP S A, SAMAD T. Imputation of missing data in industrial databases[J]. Applied intelligence, 1999, 11(3): 259–275.

[12] SONG Qinhao, SHEPPERD M, CHEN Xiangru, 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.

[13] DONDERS A R T, VAN DER HEIJDEN G J M G, STIJNEN T, et al. Review: a gentle introduction to imputation of missing values[J]. Journal of clinical epidemiology, 2006, 59(10): 1087–1091.

[14] TROYANSKAYA O, CANTOR M, SHERLOCK G, et al. Missing value estimation methods for DNA microarrays[J]. Bioinformatics, 2001, 17(6): 520–525.

[15] KEERIN P, KURUTACH W, BOONGOEN T. Cluster-based KNN missing value imputation for DNA microarray data[C]//Proceedings of 2012 IEEE International Conference on Systems, Man, and Cybernetics. Seoul, South Korea, 2012: 445-450.

[16] VAN BUUREN S, GROOTHUIS-OUDSHOORN K. Mice: Multivariate imputation by chained equations in R[J]. Journal of statistical software, 2011, 45(3): 75765.

[17] GEBREGZIABHER M, DESANTIS S M. Latent class based multiple imputation approach for missing categorical data[J]. Journal of statistical planning and inference, 2010, 140(11): 3252–3262.

[18] VERMUNT J K, VAN GINKEL J R, VAN DER ARK L A, et al. 9. Multiple imputation of incomplete categorical data using latent class analysis[J]. Sociological methodology, 2008, 38(1): 369–397.

[19] TOUTENBURG H. Rubin, D.B.: multiple imputation for nonresponse in surveys[J]. Statistical papers, 1990, 31(1): 180.

[20] SIM J M, KWON O, LEE K C. Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in data sets[J]. Expert systems with applications, 2016, 46: 485–493.

[21] 张铃, 张钹. M-P 神经元模型的几何意义及其应用 [J]. 软件学报, 1998, 9(5): 334–338.
ZHANG Ling, ZHANG Bo. A geometrical representation of M-P neural model and its applications[J]. Journal of software, 1998, 9(5): 334–338.

[22] 张燕平, 张铃. 机器学习理论与算法 [M]. 北京: 科学出版社, 2012: 56–66.

[23] JÖRNSTEN R, WANG Huiyu, WELSH W J, et al. DNA microarray data imputation and significance analysis of differential expression[J]. Bioinformatics, 2005, 21(22): 4155–4161.

[24] MAZUMDER R, HASTIE T, TIBSHIRANI R. Spectral regularization algorithms for learning large incomplete matrices[J]. The journal of machine learning research, 2010, 11: 2287–2322.

[25] RANJBAR M, MORADI P, AZAMI M, et al. An imputation-based matrix factorization method for improving accuracy of collaborative filtering systems[J]. Engineering applications of artificial intelligence, 2015, 46: 58–66.

作者简介：

严远亭，男，1986 年生，讲师，博士，中国人工智能学会会员，主要研究方向为机器学习、粒计算和生物信息学。主持国家自然科学基金青年项目 1 项，发表学术论文 10 余篇。

吴亚亚，男，1995 年生，硕士研究生，中国人工智能学会会员，主要研究方向为机器学习和不完整数据处理。

赵姝，女，1979 年生，教授，博士生导师，博士，中国人工智能学会粒计算与知识发现专委会委员，安徽省人工智能学会常务理事，主要研究方向为机器学习、粒计算。获得发明专利和软件著作权多项，发表学术论文 60 余篇。