

文章编号:1671-5896(2024)02-0372-06

基于迁移学习的非结构化大数据缺失值插补算法

颜远海, 杨莉云

(广州华商学院 数据科学学院, 广东 增城 511300)

摘要: 针对数字信息产生的海量、多角度的非结构化大数据, 由于外界干扰、数据结构损坏等因素造成其信息丢失问题, 提出了基于迁移学习的非结构化大数据缺失值插补算法。通过迁移学习算法, 预测非结构化大数据缺失部位, 利用朴素贝叶斯算法分类数据特征, 度量属性间权重值, 明确数据类别特征差异向量, 辨别特征差异程度。采用核回归模型对数据缺失部分实施非线性映射, 经过多项式变化编码, 描述数据的跨空间互补条件, 完成非结构化大数据缺失值插补。实验结果表明, 所提算法可以有效完成非结构化大数据缺失值插补, 具有较好的插补效果, 能提高插补精度。

关键词: 迁移学习; 非结构化大数据; 缺失值插补; 缺失值预测; 核回归函数

中图分类号: TP391

文献标识码: A

DOI:10.19292/j.cnki.jdxxp.2024.02.014

Missing Value Interpolation Algorithm of Unstructured Big Data Based on Transfer Learning

YAN Yuanhai, YANG Liyun

(College of Data Science, Guangzhou Huashang College, Zengcheng 511300, China)

Abstract: Due to the complexity of digital information, massive and multi-angle unstructured big data, and external interference, data structure damage and other factors cause its information loss, a missing value interpolation algorithm for unstructured big data based on transfer learning is proposed. Through the migration learning algorithm, the missing parts of unstructured big data are predicted, and the naive Bayesian algorithm is used to classify data features, to measure the weight value between attributes, to clarify the feature difference vector of data categories, and to identify the degree of feature difference. The kernel regression model is used to implement nonlinear mapping for the missing part of the data, and the polynomial change coding is used to describe the cross-space complementary condition of the data, completing the interpolation of the missing value of unstructured big data. The experimental results show that the proposed algorithm can effectively complete the interpolation of missing values of unstructured large data, has good interpolation effect and can improve the interpolation accuracy.

Key words: transfer learning; unstructured big data; imputation of missing values; missing value prediction; kernel regression function

0 引言

如果信息中存在不规则或不完整状态链条, 则称之为非结构化数据, 在这种数据结构中无法建立预定义数据模型, 数据的表现形式也无法通过数据库二维逻辑表呈现, 这类数据包括但不限于多种类型格

收稿日期: 2023-03-16

基金项目: 创新强校工程基金资助项目(2017KQNCX266)

作者简介: 颜远海(1985—), 男, 江西吉安人, 广州华商学院讲师, 主要从事大数据可视化技术和数据分析算法研究, (Tel)86-18924273591 (E-mail)yan85028@163.com。

式的办公文档、图片和报表等,甚至有一些音频视频文件中也会存在非结构化数据模式。当非结构化大数据出现缺失问题时,通常受某些因素的影响而没有获得应该处于规定位置上的信息,导致数据的统计结果缺少了一部分,降低了该数据所描述信息的真实性和可信度,致使数据分析结果与现实相比有较大的偏差。为解决上述问题,需要对丢失的数据进行缺失值插补,主要是在样本数据中找到缺失的对应值,并找出能合理替代原始丢失数据的数值,将其插补在原有缺失部位,使经过插补后的样本数据能保持所描述信息的真实和有效。

近年来,缺失数据插值方法不断得到深入研究,刘璐等^[1]针对不同模态的数据缺失情况,对原始的缺失数据进行多向延迟嵌入,获取数据折叠后的特征向量,逆向操作完成数据的平滑插补,虽然该方法具有一定的有效性,但其插补效果较差。胡雪等^[2]针对标记的缺失数据,利用张量模型进行数据多模态处理,根据不同阶层的数据特征因子,建立缺失特征变化融合矩阵,实现缺失数据的插补,虽然该方法具有较好的鲁棒性,但其插补精度较低。

针对上述问题,笔者提出了基于迁移学习的非结构化大数据缺失值插补算法。在迁移学习的基础上,对非结构化大数据缺陷部位和特征进行预测,从多个视角出发,建立缺失特征的核回归函数,完成非结构化大数据的缺失值插补。通过实验证明,所提算法能更准确地实现数据填补,还能通过参数间的调整变化,有效提高算法的插补精度,具有较好的插补效果。

1 基于迁移学习的非结构化大数据缺失值预测

非结构性大数据的特征可以进行模块度量,所有的度量值向量共同组建成非结构大数据的特征空间,假设数据的特征空间用 χ 表示,数据集表示为

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \in \chi, \quad (1)$$

其中 \mathbf{x}_n 表示非结构化大数据中第 n 个数据的度量值向量, n 表示训练样本数据的数量。在相同度量特征条件下,不同结构数据集对应的空间结构以及数据特征范围均不相同。为此,在加权贝叶斯算法^[3]基础上,建立迁移学习算法,保证能最大程度地分析和预测缺陷数据的特征信息。

假设针对非结构化数据的训练样本集合为

$$L = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}, \quad (2)$$

其中 y_n 表示第 n 个数据样本位置对应的数据类别。在进行缺失值预测过程中, $y_n \in (t', f')$, t' 表示数据中的缺失部分, f' 表示数据中未发生缺失的部分。假设 $T = \{t_1, \dots, t_m\}$ 表示非结构化数据的测试集合, m 表示测试数据中的样本数量,利用朴素贝叶斯算法对 t 进行分类:

$$y(t) = \arg \max P(y_n | t_m) = \arg \max \frac{P(y_n | t_m) \prod_{j=1}^k p(a_k | y_n)}{\sum P(y_n) \prod_{j=1}^k p(a_k | y_n)}, \quad (3)$$

其中 $t_m = \{a_1, \dots, a_j, \dots, a_k\}$, a_k 表示测试数据的第 k 个类别特征, k 表示类别特征的具体个数, $P(y_n)$ 、 $p(a_k | y_n)$ 、 $P(y_n | t_m)$ 分别表示数据缺失值的先验、条件和后验概率。在假设数据缺失类别特征基础上,利用朴素贝叶斯算法完成数据的定类分析,每个数据类别通过数据度量提取缺失特征,度量属性之间相对独立,此时认为每个度量属性之间具有相同的权重值^[4]。

在使用迁移学习预测数据缺失值过程中,每个样本数据属性为

$$\mathbf{x}_i = \{a_{i1}, a_{i2}, \dots, a_{ik}\}, \quad (4)$$

其中 a_{ik} 表示样本数据 i 对应的第 k 个类别特征。为精确获取到训练数据和测试数据集之间的相似程度,计算二者在对应类别特征上的差异情况,假定数据的属性特征向量共有5个对应元素,分别为最小、最大、平均、中位值以及标准差,如下:

$$\mathbf{A}_k = \{a_{ik} - \min_k, a_{ik} - \max_k, a_{ik} - \text{mean}_k, a_{ik} - \text{median}_k, a_{ik} - \text{std}_k\}, \quad (5)$$

其中 \mathbf{A}_k 表示第 k 个类别特征的属性向量^[5],计算训练数据和测试数据集之间的类别特征向量集合,如下:

$$A_S - T_i = \{A_1, \dots, A_k\}, \quad (6)$$

$$A_S - T_n = \{A_1, \dots, A_k\}. \quad (7)$$

通过分析数据并实施特征距离度量, 辨别数据间的特征差异程度, 距离越大表明数据间的特征差异越大, 通过欧氏距离^[6]计算上述所提到的数据特征差异, 建立数据的类别特征差异向量:

$$D = \{d_1, \dots, d_k\}, \quad (8)$$

其中 d_k 表示第 k 个类别特征的具体差异。

每个样本数据 $x_i = \{a_{i1}, a_{i2}, \dots, a_{ik}\}$ 中包含的数据特征差异可通过

$$D_{ik} = \sqrt{\frac{(a_{ik} - a_{1k})^2 + \dots + (a_{ik} - a_{mk})^2}{n}} \quad (9)$$

计算得到。其中 D_{ik} 表示样本数据 i 的第 k 个类别特征与测试数据集类别特征之间的差异, a_{mk} 表示测试数据 m 所对应的第 k 个类别特征。假设每个类别特征在测试数据中的作用相同, 通过对比 D_{ij} 和 D 之间相对应的差异值 d_k 完成相似度计算, 相似特征的数量为

$$s_i = \sum_{k=1}^i d_k(a_{ik}). \quad (10)$$

2 多视角下非结构大数据互补条件约束

利用核回归模型完成针对大数据的多视角非线性回归^[7], 进行数据缺失视角的插补约束, 利用非线性映射函数 φ 对缺失大数据进行嵌入, 将其融合到相对高维的特征空间 H 中, 在线性学习器的辅助下完成对数据高维特征空间的分析和处理。处理过程中较为重要的环节是核技巧, 在输入的样本数据所在特征空间中, 将点积运算替换成对应的函数变化式 $k(x^{(i)}, y^{(j)})$, 这里的函数变化式就是数据对应的核函数, 使数据缺失部分转换到高维特征空间^[8], 利用参数变化实现过拟合的控制约束, 缺失大数据的正定核 l 可通过非线性映射函数 φ 在空间 H 中完成点积表达:

$$\begin{cases} \varphi: \kappa \rightarrow H, \\ x \rightarrow \varphi(x), \\ l(x^{(i)}, y^{(j)}) = \langle \varphi(x^{(i)}), \varphi(y^{(j)}) \rangle, \end{cases} \quad (11)$$

其中 κ 表示数据的原始特征空间, \langle, \rangle 表示向量点积符号。

利用高斯核函数完成多项式变化编码:

$$l(x^{(i)}, y^{(j)}) = \exp(-\|x^{(i)}, x^{(j)}\|^2 / 2\sigma^2), \quad (12)$$

其中 σ 表示高斯核函数的编码条件系数。

多项式编码的数据核为

$$l(x^{(i)}, y^{(j)}) = (a \langle \varphi(x^{(i)}), \varphi(y^{(j)}) \rangle + b)^d, \quad (13)$$

其中 d 表示特征向量, a, b 表示编码数据集合的权重系数。

数据的 Sigmoid 核为

$$l(x^{(i)}, y^{(j)}) = \tanh(k(x^{(i)}, x^{(j)}) + v). \quad (14)$$

在核函数的训练矩阵中, 定义训练数据 $x^{(1)}, \dots, x^{(n)} \in \kappa$ 和对应的核函数 $k(x^{(i)}, y^{(j)})$, 此时 $n \times n$ 的核矩阵为

$$K_{ij} = k(x^{(i)}, y^{(j)}). \quad (15)$$

当数据视角 y^c 对应的类别特征完全缺失时, 可通过将不同视角下的特征进行串联融合, 在单视角基础上完成相应的数据缺失插补, 再进一步从多个视角下针对各个数据间存在的类别特征完成互补, 考虑到在不同视角下的大数据具备横跨异构空间的能力, 引领多视角下的缺失数据完成关系对应, 在核函数基础上建立多视角下的数据非线性关系模型, 主要描述了数据在跨空间中的数据互补条件^[9]。

针对存在缺陷值的多视角大数据 $\{X, Y\}$, 通过对视角 X 下的缺陷数据 $X = [x^{(1)}, \dots, x^{(n)}]$ 进行对应的非线性函数 $\varphi(x)$ 映射, 在相对高维的数据特征空间 H 中完成, 通过线性回归分析方法完成对数据 $\{\varphi(X), Y\}$ 的建模, 其中 $\varphi(X)$ 为

$$\varphi(X) = [\varphi x^{(1)}, \dots, \varphi x^{(N)}]. \quad (16)$$

建模为

$$\min_{w_g}(w) = \frac{\lambda}{n}w + \frac{1}{n} \sum_{j=1}^n (y^{(j)}[d] - w^T \varphi(X^{(j)}))^2, \quad (17)$$

其中 $y^{(j)}[d]$ 表示 j 缺失数据的类别特征向量 d 的高维对应变量值, $d=1,2,\dots,s$, w 表示数据线性回归模型的回归系数^[10], λ 表示模型的回归正则调整参数。

计算模型的回归系数 w 得到最优解, 通过对缺失数据的线性训练^[11]得到:

$$w^* = \sum_{j=1}^n \beta_j \varphi(x^{(j)}) = \varphi(X)\beta, \quad (18)$$

其中 $\beta \in R^{n \times 1}$ 表示缺失数据的训练权重向量, w^* 表示最优参数 φ , 将式(17)与式(18)相结合可得:

$$\min_{\beta} g(\beta) = \frac{\lambda}{n} \sum_{n=1}^n \sum_{j=1}^n \beta_i \beta_j \langle \varphi(x^{(i)}), \varphi(x^{(j)}) \rangle + \frac{1}{n} \sum_{n=1}^n (y^{(i)}[d] - \sum_{j=1}^n \beta_j \langle \varphi(x^{(i)}), \varphi(x^{(j)}) \rangle)^2. \quad (19)$$

利用核技巧有效地简化公式可得:

$$\begin{aligned} \min_{\beta} g(\beta) &= \frac{\lambda}{n} \sum_{n=1}^n \sum_{j=1}^n \beta_i \beta_j l(x^{(i)}, y^{(j)}) + \frac{1}{n} \sum_{n=1}^n (y^{(i)}[d] - \sum_{j=1}^n \beta_j l(x^{(i)}, y^{(j)}))^2 = \\ &= \frac{\lambda}{n} \beta^T K \beta + \frac{1}{n} (\beta^T K^T K \beta - 2 \beta^T K^T (y[d])^T + y[d] y[d]^T), \end{aligned} \quad (20)$$

其中 $K \in R^{n \times n}$ 表示缺失数据集 X 对应的核矩阵, $y[d] \in R^{1 \times n}$ 表示缺失大数据集合 Y 对应的第 d 维特征变量, 则得:

$$\nabla g(\beta) = \frac{2}{n} (\lambda K^T \beta + K^T K \beta - K^T y[d]^T) = \frac{2}{n} K^T ((\lambda I + K)\beta - y[d]^T). \quad (21)$$

通过设定 $\nabla g(\beta) = 0$ 计算得出参数 β 的最优解:

$$\beta = (\lambda I + K)^{-1} (y[d]^T)^T. \quad (22)$$

3 非结构化大数据缺失值插补算法

基于迁移学习完成了对非结构性大数据的缺失值预测后, 并设置了多视角下的缺失互补条件约束, 最后建立分数热卡插补法, 对缺失数据进行分层均值建模^[12], 实现缺失值插补。假设数据的个体总数 U 共由 G 个数据分组组成, 则此时可将 G 个数据分组作为缺失数据的插补分层, 在单独的层中, 有限的数数据个体中包含了大量的特征元素, 且每个特征元素均符合随机变量下的独立分布状态, 假设数据的特征分布均值为 μ_g , 函数计算的方差为 σ_g^2 , 此时插补层中包含的元素分布为

$$y_i^{i.i.} = (\mu_g, \sigma_g^2), \quad (23)$$

其中 $y_i^{i.i.}$ 表示数据缺失特征的独立分布状态。

热卡插补法主要通过缺失项和缺失赋值元素实现缺失值插补, 其中分布状态表示为 $d = \{d_{ij}: i \in A_r, j \in A_m\}$, d_{ij} 表示数据的缺失项 y_j 对应到的赋值元素, 其中包含特征分布的插补机制。

另外的要素为与数据缺失项相对应的赋值元素权重, 利用 w_{ij}^* 作为特征元素的权重分数, 描述缺失元素的特征条件, 需要通过数据的缺失元素完成加权平均值的插补值:

$$y_{ij} = \sum_{i \in A_R} w_{ij}^* y_i^{i.i.}. \quad (24)$$

特征元素的插补值可有效地完成对缺失数据的填补, 利用 A_{rg} ($A_{rg} > 0$) 代表特征元素层 g 中包含的回答部分, A_{mg} 代表特征元素层 g 中包含的无回答部分, A_{Dj} 表示特征元素层 g 中的补充元素, j 属于 A_{mg} , A_{Dj} 表示 A_{rg} 的对应数据子集, 在数据特征的分层元素均值模型下, 针对数据的线性热卡插补计算如下:

$$\hat{\theta}_n^l = \sum_{g=1}^G \left(\sum_{i \in A_{rg}} w_i y_i^{i.i.} + \sum_{j \in A_{mg}} w_j \sum_{i \in A_{Dj}} w_{ij}^* y_i^{i.i.} \right) y_{ij}. \quad (25)$$

所有缺失数据的赋值元素通过权重调配, 得到对应的权重总和:

$$w_{ij}^{**} = w_j \hat{\theta}_n^l, \quad (26)$$

其中 w_{ij}^{**} 表示总权重, 此时得到的数据估计值为数据的分数热卡插补结果。

4 实验分析

为验证所提的基于迁移学习的非结构化大数据缺失值插补算法实际性能, 在缺失数据样本集合中随机抽取 20% 作为测试集合, 其余的 80% 作为缺失数据插补算法的训练样本集合。

在进行数据插补前, 由于所提算法中受到针对缺失数据的核回归函数影响, 其中包含的回归正则调整参数 λ 、核函数编码条件系数 σ 为主要影响因素, 为使后续数据插补的过程中能获得更真实有效的性能结果, 需要对参数条件进行实验。以缺失数据集合为例, 基于核回归函数完成缺失插补的实验中, 当参数 $\sigma=5$ 时, 参数 λ 对数据的缺失值插补影响如图 1 所示, 当参数 $\lambda=0.1$ 时, 参数 σ 对插补性能的影响如图 2 所示。

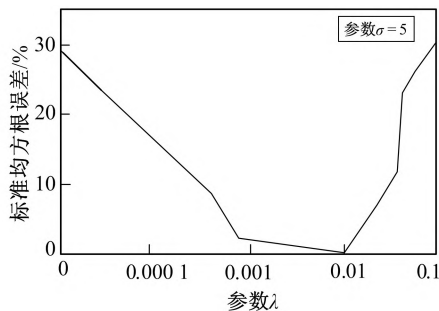


图1 参数 λ 对插补性能影响

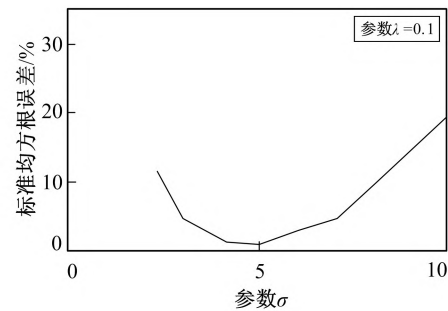


图2 参数 σ 对插补性能影响

Fig.1 Effect of parameter λ on interpolation performance Fig.2 Effect of parameter σ on interpolation performance

从图1、图2中可看出, 当 λ 取值为 0.01 时, 可得到最优的数据缺失值插补效果, 而当 λ 的取值过大或过小时, 都存在很大的概率会使数据缺失值的核回归性能降低; 当 σ 的取值维持在 5 时, 可得到数据缺失插补的最优效果, 若是过大或过小同样会降低算法的插补性能。

使用上述测试出的最优参数, 完成对不同类型数据的缺失插补实验。实验使用两种类型的数据集合, 其中一种仅包含文字信息, 另外一种仅包含图像信息, 随机抽取 70% 的数据样本作为插补训练集合, 其余的 30% 作为数据插补测试集合, 并使用多模态插补算法(文献[1]方法)、数据张量插补算法(文献[2]方法)作为对比方法。文字非结构化大数据集合下不同方法的插补性能如图 3 所示。

从图 3 可看出, 所提算法获得的插补效果最好, 标准均方根误差与其他两种方法比最小, 这是因为所提算法基于迁移学习的缺失值预测, 在核回归计算基础上, 完成多视角数据的缺失插补, 能有效地提高文字非结构化大数据的插补精度。图像非结构化大数据集合下不同方法插补性能如图 4 所示。

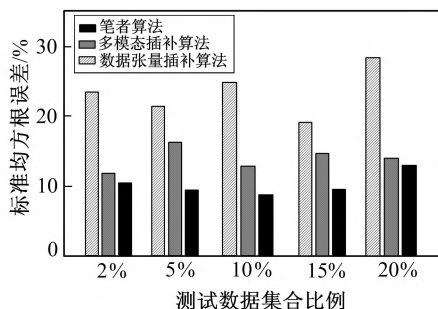


图3 文档数据集插补的标准均方根误差

Fig.3 Standard root mean square error of document dataset interpolation

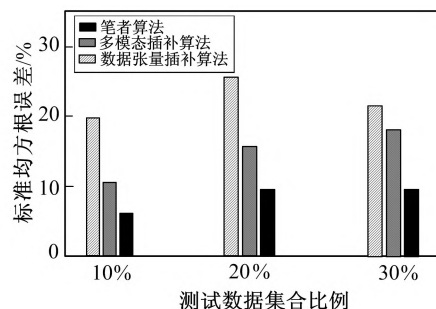


图4 图像数据集插补的标准均方根误差

Fig.4 Standard root mean square error of image dataset interpolation

从图 4 可看出, 所提算法具有较好的插补效果, 在预测数据缺失位置及特征的前提下, 能更好地完成对图像非结构化大数据多个视角的相关性捕捉, 增强数据插补的性能, 降低数据插补误差。

5 结 语

针对存在缺失值的非结构大数据,笔者提出了一种基于迁移学习的非结构化大数据缺失值插补算法。通过迁移学习完成对非结构化大数据缺失位置及特征的预测,在核回归函数的多角度下,实现数据缺失特征的重构,构造多视角之间的非线性回归模型,捕捉数据之间的相互补充性,最终有效地完成缺失数据的插补。经实验证明,所提算法的插补效果较好,能有效提高插补精度。

参考文献:

- [1] 刘璐,杜若画,吴强,等. 一种基于张量分解的医学数据缺失模态的补全算法[J]. 数据采集与处理, 2021, 36(1): 45-52.
LIU J, DU R H, WU Q, et al. A Complete Algorithm for Missing Modalities of Medical Data Based on Tensor De-Composition [J]. Journal of Data Acquisition & Processing, 2021, 36(1): 45-52.
- [2] 胡雪,彭敦陆. 张量表达下的多模态交通缺失数据补全算法[J]. 小型微型计算机系统, 2021, 42(1): 105-110.
HU X, PENG D L. Imputation Algorithm for Multimodal Missing Traffic Data Based on Tensor [J]. Journal of Chinese Computer Systems, 2021, 42(1): 105-110.
- [3] 邱德俊,周洋,仲静文,等. 基于空间邻近点与极限学习机的大坝位移缺失数据补齐[J]. 水力发电, 2021, 47(12): 98-101.
QIU D J, ZHOU Y, ZHONG J W, et al. Missing Data Complement of Dam Displacement Based on Spatial Proximity Points and Extreme Learning Machine [J]. Water Power, 2021, 47(12): 98-101.
- [4] 方匡南,谢邦昌. 基于聚类关联规则的缺失数据处理研究[J]. 统计研究, 2011, 28(2): 87-92.
FANG K N, XIE B C. Research on Dealing with Missing Data Based on Clustering and Association Rule [J]. Statistical Research, 2011, 28(2): 87-92.
- [5] 崔阳阳,赵洪山,曲岳哈,等. 基于残差U型网络的低压台区电力缺失数据补全方法[J]. 电力系统自动化, 2022, 46(9): 83-90.
CUI Y Y, ZHAO H S, QU Y H, et al. Residual U-Net Based Complementation Method for Missing Electricity Data of Low-Voltage Stations [J]. Automation of Electric Power Systems, 2022, 46(9): 83-90.
- [6] 冯磊,王石刚,梁庆华. 基于GAKNN方法的配电站时间序列缺失数据补全方法[J]. 电力自动化设备, 2021, 41(12): 187-192.
FENG L, WANG S G, LIANG Q H. Completion Method for Missing Time Series Data of Distribution Station Based on GAKNN Method [J]. Electric Power Automation Equipment, 2021, 41(12): 187-192.
- [7] 谢佳鑫,俞卫琴. 张量加权 Schatten 范数交通数据补全估计[J]. 公路交通科技, 2021, 38(12): 122-130.
XIE J X, YU W Q. Estimation on Tensor Weighted Schatten Norm for Missing Traffic Data Completion [J]. Journal of Highway and Transportation Research and Development, 2021, 38(12): 122-130.
- [8] 陈传毅,戴卫军. 基于贝叶斯网的高维数据隐藏模式挖掘[J]. 计算机仿真, 2021, 38(1): 287-290, 349.
CHEN C Y, DAI W J. High Dimensional Data Hiding Pattern Mining Based on Bayesian Network [J]. Computer Simulation, 2021, 38(1): 287-290, 349.
- [9] 李培冠,於志勇,黄昉苑. 基于稀疏表示的电力负荷数据补全[J]. 计算机科学, 2021, 48(2): 128-133.
LI P G, YU Z Y, HUANG F W. Power Load Data Completion Based on Sparse Representation [J]. Computer Science, 2021, 48(2): 128-133.
- [10] 刘云,郑文凤,张轶. 对称加权算法对数据矩阵补全的优化研究[J]. 四川大学学报(自然科学版), 2021, 58(4): 79-86.
LIU Y, ZHENG W F, ZHANG Y. Optimization of Data Matrix Completion by Symmetric Weighting Algorithm [J]. Journal of Sichuan University (Natural Science Edition), 2021, 58(4): 79-86.
- [11] 雷明阳,陈静杰,欧晓勇,等. 基于张量分解的智能电表电压数据缺失填补算法[J]. 电网与清洁能源, 2021, 37(12): 8-15.
LEI M Y, CHEN J J, OU X Y, et al. A Missing Voltage Data Imputation Algorithm for Smart Meters Based on Tensor Decomposition [J]. Advances of Power System & Hydroelectric Engineering, 2021, 37(12): 8-15.
- [12] 郭毅博,牛猛,王海迪,等. 基于生成对抗网络的飞机燃油数据缺失值填充方法[J]. 浙江大学学报(理学版), 2021, 48(4): 402-409.
GUO Y B, NIU M, WANG H D, et al. An Aircraft Fuel Data Missing Value Filling Method with Generative Adversarial Network [J]. Journal of Zhejiang University (Science Edition), 2021, 48(4): 402-409.

(责任编辑:刘东亮)