doi：10.13413/j.cnki.jdxblxb.2024224

基于泛化中心聚类的时间序列缺失数据填补方法

于艳朋，惠向晖
（河南农业大学 信息与管理科学学院(软件学院)，郑州 450046）

摘要：针对填补时间序列中的缺失值通常依赖于已有数据的预测，由于时间序列的复杂性和不确定性导致预测结果常存在误差的问题，为保证数据填补效果，提出一种基于泛化中心聚类的时间序列缺失数据填补方法. 首先，计算对象与类之间、类与类之间的距离，量化数据点与聚类中心之间的相对位置关系，得到数据间的空间关系. 其次，利用信息瓶颈算法对空间中的泛化中心进行聚类处理，将含有缺失数据的时间序列数据集划分到同一类中. 最后，计算簇半径，对泛化中心聚类后产生的离群点数据再次进行可用、弱可用随机损坏数据划分，设置波动阈值，将位于波动阈值内的随机损坏数据与聚类中统一属性值进行字符串对比，实现时间序列缺失数据填补. 实验结果表明，该方法在聚类过程中有较高的标准化互信息和命中率，在缺失数据填补时，可保证数据补齐率在 80％以上，说明该方法可有效改善时间序列数据的完整性.

关键词：泛化中心聚类；时间序列；缺失数据填补；信息瓶颈；随机损坏数据；补齐率
中图分类号：TP391 **文献标志码**：A **文章编号**：1671-5489(2025)04-1137-06

Missing Data Filling Method in Time Series Based on Generalized Center Clustering

YU Yanpeng, HUI Xianghui
(*College of Information and Management Science* (*College of Software*), *Henan Agricultural University*, *Zhengzhou* 450046, *China*)

Abstract: Aiming at the problem that the filling of missing values in time series usually relied on the predictions of existing data, and the complexity and uncertainty of time series often led to errors in the prediction results. In order to ensure the effectiveness of data filling, we proposed a time series missing data filling method based on generalized center clustering. Firstly, we calculated the distance between objects and classes, as well as between classes, quantified the relative positional relationship between data points and cluster centers, and obtained the spatial relationship between data. Secondly, we used information bottleneck algorithms to cluster the generalization centers in space, dividing time series datasets containing missing data into the same class. Finally, we calculated the cluster radius, divided the outlier data generated by the generalized center clustering into usable and weakly usable randomly damaged data, set a fluctuation threshold, and compared the randomly damaged data within

收稿日期：2024-06-04.

第一作者简介：于艳朋(1987—)，男，汉族，硕士，讲师，从事计算机应用、计算机教育教学、计算机理论、软件工程、数据科学及大数据技术的研究，E-mail：yuyanpeng19870@163.com.

基金项目：国家自然科学基金(批准号：62176239).

the fluctuation threshold with a string of the unified attribute values in the cluster, achieving the missing data filling in the time series. The experimental results show that this method has high standardized mutual information and hit rate in the clustering process, and can ensure a data replenishment rate of over 80% when filling in missing data, indicating that this method can effectively improve the integrity of time series data.

Keywords: generalized center clustering; time series; missing data filling; information bottleneck; randomly damaged data; replenishment rate

随着时间序列数据在经济学、金融学、气象学等多个领域的广泛应用，数据的准确性和完整性对分析和预测时间相关现象和趋势至关重要[1]. 但在实际应用中时间序列数据常面临缺失问题[2-4]，因此研究时间序列缺失数据的填补方法具有重要意义. 其旨在通过合适的数学和统计方法，有效恢复缺失数据，提高数据分析的质量和效率，对推动相关领域的研究和决策具有重要意义.

目前，对缺失数据填补方法的研究已取得了很多成果. 例如：乔非等[5]研究了面向多维特性数据的缺失值检测及填补方法，该方法检测了多维数据的缺失程度，在不同缺失程度下设计了不同填补方法，但该方法在进行缺失数据填补时无法保证数据补齐率，导致填补结果不完善；任兵等[6]研究了基于压缩感知的相关性数据填补方法，该方法将填补问题转化为压缩感知框架下的稀疏向量恢复问题，通过快速迭代加权阈值算法实现缺失数据的填补，虽然在填补过程中效率较高，但其并不适用于时间序列数据；卢继哲等[7]通过神经网络模型预测缺失内容，并对其进行填补，虽然缺失数据预测的结果较精准，但其在聚类过程中无法保证较高的标准化互信息，从而无法保证后续填补质量；Sun等[8]研究了基于缺失率和异常度测量的不完全数据处理方法，该方法对不同缺失比率异常数据进行检测，并通过填补方法对其填充，该方法虽然能实现不同缺失比率异常数据的精准检测，但填补后仍存在大量缺失数据. 为保证缺失数据填补质量，本文提出一种基于泛化中心聚类的时间序列缺失数据填补方法.

1 时间序列缺失数据填补

通过泛化中心聚类可将具有相似时间模式或特征的数据点聚集在一起，从而利用这些相似数据的信息更准确地估计和填补缺失值[9-10]. 这种方法有助于捕捉时间序列数据的内在结构，提高缺失数据填补的准确性和可靠性.

1.1 泛化中心距离计算

计算泛化中心距离的意义在于衡量不同数据点在特征空间中的相对位置关系，从而评估不同数据之间的相似性或差异性. 对象(数据点)与类(聚类)之间距离的计算公式为

$$d(\boldsymbol{x},\boldsymbol{O})=\sqrt{\sum_{i=1}^{n}(x_i-O_i)^2}, \tag{1}$$

其中 $\boldsymbol{x}$ 为对象(数据点)的特征向量，$\boldsymbol{O}$ 为泛化中心的特征向量，n 表示特征的数量. 通常情况下，距离 d 越小，说明对象与该类中的其他对象越类似，设置距离上限为 U，若某一对象的距离 $d\leqslant U$，则将该对象添加到该类中，否则构建一个新类[11].

当计算完对象与类之间的距离后，需计算类与类之间的距离，分别设两个类为 O_1,O_2，两者的泛化中心分别为 O_{O1} 和 O_{O2}，则这两个类之间距离的计算公式为

$$d(O_1,O_2)=d(O_{O1},O_{O2})\times(\sqrt{|O_1|}+\sqrt{|O_2|})/2, \tag{2}$$

其中 $|O_1|,|O_2|$ 分别表示类 O_1,O_2 中存在的对象数量，若某一类中仅存在一个对象，则可将该对象设为泛化中心. 通过类间距离 $d(O_1,O_2)$ 可以评估不同类之间的差异性，$d(O_1,O_2)$ 越小两个类之间越相似，若该值越大，则两个类之间的差异就越明显.

1.2 基于信息瓶颈算法的泛化中心聚类

通过上述计算得到了对象与类之间、类与类之间的距离，量化了数据点与聚类中心之间的相对位

置关系，得到了数据间的空间关系[12-13]，然后可利用信息瓶颈(information bottleneck，IB)算法对空间中的泛化中心进行聚类处理. 信息瓶颈算法通过限制信息传递的瓶颈，能在保持数据聚类质量的同时，减少不必要的信息损失[14]，从而实现数据的有效压缩和聚类.

设原始数据集为 $X=\{x_1,x_2,\cdots,x_N\}$，聚类中心为 $C=\{c_1,c_2,\cdots,c_K\}$，其中 K 为预设的聚类数量. 定义一个编码函数 $q(c|x)$和一个解码函数 $p(c|x)$. 使用互信息度量 X 与 C 之间的相关性：

$$I(X,C)=\sum_{x\in X}\sum_{c\in C}p(x,c)\log_2\frac{p(x,c)}{p(x)p(c)}, \tag{3}$$

其中 $p(x)$，$p(c)$分别为 X 和 C 的边缘概率分布，$p(x,c)$为先验联合分布. 使用条件熵度量 X 在给定 C 下的不确定性：

$$H(X|C)=-\frac{\sum_{x\in X}\sum_{c\in C}p(x,c)}{\log_2 p(x|c)}. \tag{4}$$

信息瓶颈的目标是最小化 $H(X|C)$的同时最大化 $I(X,C)$，可通过优化一个加权和实现：

$$\min_{q(c|x)}\{\beta I(X,C)-H(X,C)\}, \tag{5}$$

其中 β 为一个权衡参数.

利用迭代方法求解上述优化问题，迭代过程中更新编码函数 $q(c|x)$和解码函数 $p(c|x)$，以减少目标函数的值. 经过多次迭代后，可得到优化的编码函数 $q(c|x)$和解码函数 $p(c|x)$. 对于每个聚类中心 c_k，可使用解码函数生成或描述该簇的泛化中心. 这种泛化中心聚类方法不仅考虑了聚类中心的位置，还考虑了数据在聚类中心周围的分布情况，因此能提供更丰富的信息表示每个簇.

1.3　基于泛化中心聚类的缺失数据填补

通过泛化中心聚类先将含有缺失数据的时间序列数据集划分到同一类中，然后再对缺失数据进行填补. 在数据填补过程中，通常会将完全缺失和部分缺失的情况都归为缺失数据，这种方式忽视了部分缺失数据所蕴含的潜在价值. 因此，本文在处理缺失数据填补问题时，先对数据损坏程度进行细致的划分：若残留字符与原始数据毫无关联，则该数据被界定为弱可用随机损坏数据；反之，为可用随机损坏数据. 通过区分并有效利用这两种数据，可进一步提升缺失数据填补的准确性和可靠性.

基于上述数据划分结果，采用簇半径计算方法对泛化中心聚类后产生的离群点数据再次进行可用、弱可用随机损坏数据划分. 在实际划分时，给定波动阈值 ϑ，将位于波动阈值内的随机损坏数据与聚类中统一属性值进行字符串对比，波动阈值 ϑ 的计算公式为

$$\vartheta=R_c+\varpi, \tag{6}$$

其中 R_c 表示泛化中心簇半径，ϖ 表示簇头竞争半径. 簇中心点 C_m 可利用下式计算：

$$C_m=\sum_{i=1}^{N}\frac{t_{ip}}{N}-\vartheta, \tag{7}$$

其中 t_{ip} 为泛化中心任一点，N 为被选择的泛化中心点数量. 此时，簇半径可表示为

$$R_c=\sqrt{\sum_{i=1}^{N}\frac{(t_{ip}-C_m)^2}{N}}. \tag{8}$$

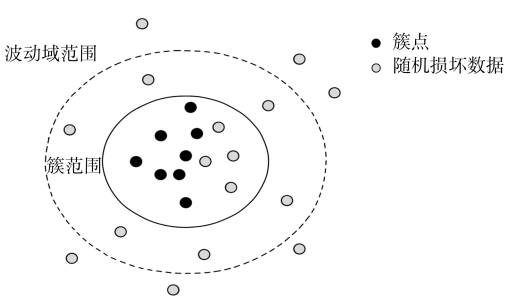

图 1　数据损坏程度划分

Fig. 1　Classification of data damage levels

通过上述计算，对泛化中心聚类后产生的离群点再次进行缺失数据划分，以此为基础，降低弱可用数据的占比. 图 1 为数据损坏程度划分.

在进行时间序列缺失数据填补过程中，若找到一个匹配的字符串，则将其记录为 1；如果未找到匹配的字符串，则记录为 0. 然后将所有字符串对比得到的数值相加，得到一个总和 b，b 值越大，说明对象之间的相似度越高. 基于这一原理，利用 b 值最大的项所对应的数据对缺失数据进行填补.

该过程不仅考虑了部分缺失数据的重要性，还提高了数据填补的准确性.

2 实验分析

为评估本文方法对时间序列缺失数据的填补能力，下面对该方法进行性能测试. 在UCI数据库中选择实验所用数据集，数据集信息列于表1.

表1 实验数据集信息

Table 1 Information of experimental dataset

数据集	样本数/个	属性维数	缺失数/个	数据集	样本数/个	属性维数	缺失数/个
Car evalution	1 878	6	752	Boston housing data	505	13	313
Mushroom	8 311	22	392	Credit approval	762	12	182
Abalone	4 721	8	318				

标准化互信息(NMI)可评估聚类结果与数据实际类标签之间的相似性，本文利用NMI评估聚类质量：

$$\mathrm{NMI}(S,L)=\frac{E(S,L)}{\sqrt{H(S)H(L)}}\times 100\%, \tag{9}$$

其中S表示目标聚类结果，L表示数据实际类标签，$E(S,L)$表示S与L之间的互信息，$H(S)$,$H(L)$分别表示S,L的信息量. NMI值越大，说明聚类结果与实际情况越相符. 利用命中率(HR)指标评估本文方法的聚类质量：

$$\mathrm{HR}=\frac{h_{\mathrm{its}}}{\psi}, \tag{10}$$

其中ψ表示缺失数据总数，h_{its}表示成功聚类数据. 利用补齐率(CR)评估本文方法的缺失数据填补能力：

$$\mathrm{CR}=\frac{C/n+F/n}{M\times\zeta}, \tag{11}$$

其中F为错误填补的数据总量，ζ为缺失数据比率，M为数据总量.

选取文献[5]方法、文献[6]方法和文献[7]方法与本文方法进行对比，不同方法的标准化互信息对比结果列于表2. 由表2可见，文献[5]方法的标准化互信息相对较小，文献[6]方法和文献[7]方法的标准化互信息虽然高于文献[5]方法，但并未超过50%，而本文方法在聚类过程中可提供较高的标准化互信息. 可见，相比于其他方法，本文方法具有良好的聚类能力.

表2 不同方法的标准化互信息对比结果

Table 2 Comparative results of standardized mutual information of different methods %

数据集	文献[5]方法	文献[6]方法	文献[7]方法	本文方法
Car evalution	33.32±2.45	39.86±1.53	34.56±1.87	52.43±1.42
Mushroom	31.64±2.42	36.86±1.67	33.76±2.53	51.67±2.26
Abalone	36.54±1.63	42.54±2.42	40.96±3.26	50.63±0.75
Boston housing data	37.85±1.89	41.75±1.54	37.75±2.64	52.45±1.53
Credit approval	35.42±0.67	44.43±2.64	36.75±2.16	53.67±1.86

选取数据集Abalone和Mushroom，用本文方法对该数据集进行聚类和缺失数据填补，分析本文方法在处理该数据集不同缺失比率时的命中率和补齐率，分析结果如图2所示. 由图2可见，当数据集中缺失数据比率逐渐增大时，本文方法对数据处理时的命中率和补齐率也随之降低，但在本文方法处理下，命中率始终保持在70%以上，缺失数据补齐率始终保持在80%以上，可见本文方法具有良好的缺失数据填补能力. 这是因为本文方法在泛化中心聚类后，对产生的离群点数据进行了再次处理，将其划分为可用、弱可用随机损坏数据，该策略能更精细地处理数据集中的异常值，提高数据填补的精度.

利用本文方法对不同数据集进行缺失数据填补，并分析用该方法进行填补后的数据缺失数，以此评估该方法的数据填补能力，分析结果列于表3.

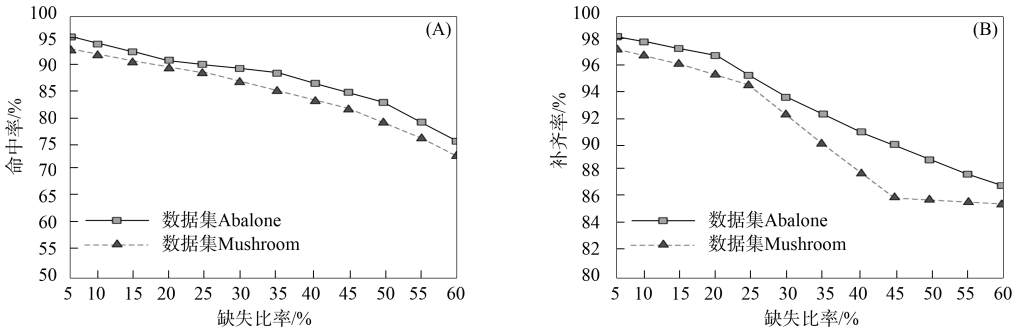

图 2　缺失数据命中率(A)和补齐率(B)测试结果

Fig. 2　Test results of missing data hit rate (A) and replenishment rate (B)

表 3　缺失数据填补能力测试结果

Table 3　Test results of missing data filling ability

数据集	样本数/个	缺失数/个	经本文方法填补后缺失数/个
Car evalution	1 878	752	32
Mushroom	8 311	392	72
Abalone	4 721	318	66
Boston housing data	505	313	45
Credit approval	762	182	26

由表 3 可见，本文方法将泛化中心聚类方法引入到时间序列缺失数据填补中，通过计算数据点与聚类中心之间的相对位置关系，得到数据间的空间关系，能更准确地捕捉时间序列数据的内在结构和特征. 因此，用本文方法对数据进行填补后，不同数据集中的缺失数据数量明显下降，其中，数据集 Credit approval 在填补后缺失数据数量仅为 26 个，可见本文方法在进行缺失数据填补时可有效保证填补质量.

综上所述，针对填补时间序列中的缺失值通常依赖于已有数据的预测，由于时间序列的复杂性和不确定性导致预测结果常存在误差的问题，本文提出了一种基于泛化中心聚类的时间序列缺失数据填补方法. 首先，该方法通过引入信息瓶颈算法对时间序列数据进行泛化中心聚类，精确捕捉数据的内在结构和特征. 其次，针对聚类后产生的离群点数据，进一步采用波动阈值与字符串对比的策略进行精细处理，实现了缺失数据的准确填补. 最后，通过量化指标如标准化互信息、命中率及数据补齐率，对本文方法的填补效果进行了全面评估，结果表明，本文方法的命中率始终保持在 70%以上，缺失数据补齐率始终保持在 80%以上，有效提升了缺失数据填补效果.

参　考　文　献

[1] 关李晶，何洁帆，张立勇，等. 基于单输出子网迭代学习的缺失值填补方法 [J]. 大连理工大学学报，2022，62(4)：427-432. (GUAN L J，HE J F，ZHANG L Y，et al. Missing Value Imputation Method Based on Single Output Sub-network with Iterative Learning [J]. Journal of Dalian University of Technology，2022，62(4)：427-432.)

[2] 陈俊扬，戴志江，李雪亮，等. 基于强化学习的多变量时序数据缺失值补全方法 [J]. 中国科技论文，2023，18(11)：1205-1212. (CHEN J Y，DAI Z J，LI X L，et al. Reinforcement Learning Based Missing Value Completion Method for Multivariate Time Series Data [J]. China Sciencepaper，2023，18(11)：1205-1212.)

[3] 邓明星，欧阳含笑，钱枫，等. 基于改进 LSTM 的重型柴油车远程监测 NO_x 浓度缺失数据填补 [J]. 环境科学学报，2023，43(11)：245-257. (DENG M X，OUYANG H X，QIAN F，et al. Filling in Missing NO_x Concentration Data for Remote Monitoring of Heavy-Duty Diesel Vehicles Based on Improved LSTM [J]. Acta Scientiae Circumstantiae，2023，43(11)：245-257.)

[4] 刘兵，郑承利. 基于 EMD 特征提取的高频面板数据自适应聚类方法 [J]. 统计与决策，2022，38(10)：16-20. (LIU B，ZHENG C L. Adaptive Clustering Method for High Frequency Panel Data Based on EMD Feature

Extraction [J]. Statistics & Decision, 2022, 38(10): 16-20.)

[5] 乔非, 翟晓东, 王巧玲. 面向多维特性数据的缺失值检测及填补方法对比 [J]. 同济大学学报(自然科学版), 2023, 51(12): 1972-1982. (QIAO F, ZHAI X D, WANG Q L. Comparison of Imputation Methods Based on Missing Value Detection for Multidimensional Feature Data [J]. Journal of Tongji University (Natural Science), 2023, 51(12): 1972-1982.)

[6] 任兵, 郭艳, 李宁, 等. 基于压缩感知的相关性数据填补方法 [J]. 计算机科学, 2023, 50(7): 82-88. (REN B, GUO Y, LI N, et al. Method for Correlation Data Imputation Based on Compressed Sensing [J]. Computer Science, 2023, 50(7): 82-88.)

[7] 卢继哲, 刘宣, 唐悦, 等. 基于聚类和LSTM的电力分钟冻结数据缺失值填充方法 [J]. 控制工程, 2022, 29(4): 611-616. (LU J Z, LIU X, TANG Y, et al. Missing Value Treatment for Minute Freezing Data of Electricity Based on Clustering and LSTM [J]. Control Engineering of China, 2022, 29(4): 611-616.)

[8] SUN Z G, GAO M M, JIANG A P, et al. Incomplete Data Processing Method Based on the Measurement of Missing Rate and Abnormal Degree: Take the Loose Particle Localization Data Set as an Example [J]. Expert Systems with Applications, 2023, 216(4): 119411-1-119411-22.

[9] 肖钊, 邓杰文, 刘晓明, 等. 基于运行规律和TICC算法的风电SCADA高维时序数据聚类方法 [J]. 机械工程学报, 2022, 58(23): 196-207. (XIAO Z, DENG J W, LIU X M, et al. Clustering Method of High-Dimensional Time Series SCADA Data from Wind Turbines Based on Operational Laws and TICC Algorithm [J]. Journal of Mechanical Engineering, 2022, 58(23): 196-207.)

[10] 李建华, 朱泽阳, 徐礼胜, 等. 基于深度嵌入聚类的ICU患者生理数据缺失插补 [J]. 东北大学学报(自然科学版), 2022, 43(5): 639-645. (LI J H, ZHU Z Y, XU L S, et al. Interpolation of Missing Physiological Data of ICU Patients Based on Deep Embedded Clustering [J]. Journal of Northeastern University (Natural Science), 2022, 43(5): 639-645.)

[11] 刘恒孜, 吕宁, 姜侯, 等. 基于DCT-PLS算法的MODIS LST缺值填补方法研究 [J]. 地球信息科学学报, 2022, 24(2): 378-390. (LIU H Z, LÜ N, JIANG H, et al. Research on Gaps Filling of MODIS LST Based on DCT-PLS [J]. Journal of Geo-information Science, 2022, 24(2): 378-390.)

[12] 赵林锁, 陈泽, 丁琳琳, 等. 基于RELM的时间序列数据加权集成分类方法 [J]. 计算机工程与科学, 2022, 44(3): 545-553. (ZHAO L S, CHEN Z, DING L L, et al. A Weighted Ensemble Classification Method for Time Series Data Based on Regularized Extreme Learning Machine [J]. Computer Engineering & Science, 2022, 44(3): 545-553.)

[13] LIU S S, HU R, WU J F, et al. Research on Data Classification and Feature Fusion Method of Cancer Nuclei Image Based on Deep Learning [J]. International Journal of Imaging Systems and Technology, 2022, 32(3): 969-981.

[14] 古险峰, 汤永利. 基于群体智能算法的混合属性大数据聚类仿真 [J]. 计算机仿真, 2023, 40(9): 458-461. (GU X F, TANG Y L. Simulation of Mixed Attribute Big Data Clustering Based on Swarm Intelligence Algorithm [J]. Computer Simulation, 2023, 40(9): 458-461.)

(责任编辑: 韩 啸)