基于GAKNN方法的配电站时间序列缺失数据补全方法

冯 磊，王石刚，梁庆华
（上海交通大学 机械与动力工程学院，上海 200240）

摘要：配电站时间序列数据从采集、传输到存储的过程中可能出现数据记录缺失的情况，在一定程度上影响高层级的数据分析及处理。针对这一问题，提出一种基于灰色自适应*K*-最近邻（GAKNN）方法的缺失数据补全方法。首先构建时间序列特征，然后在朴素*K*-最近邻（KNN）方法的基础上设置阈值筛选最近邻点，并结合灰色关联系数计算近邻点权重系数，最终依次补全缺失数据。以江苏省某市的电力数据样本进行实验，结果表明与其他方法进行对比，基于GAKNN方法的缺失数据补全方法的结果更好，并且补全后的样本在深度学习预测中具有更低的误差。

关键词：配电站数据；时间序列；数据补全；灰色自适应*K*-最近邻
中图分类号：TM 93　　**文献标志码**：A　　DOI：10.16081/j.epae.202108004

0 引言

随着配电站逐步智能化，产生的数据量日益庞大[1]，在数据采集、传输及存储的过程中，由于传感器空采样、通信异常等不可控因素难免出现数据缺失的状况。样本的缺陷导致可用数据减少、时间关联性被破坏；数据量的减少会造成训练模型无法被完全驱动、相关模型的结果存在偏差，为后续分析带来困难[2-3]。

针对时间序列数据缺失的问题，传统的缺失数据补全（下文简称数据补全）方法主要包括均值插补、多项式插补法等[4]。均值插补和多项式插补法直接对时间序列数据进行拟合，未考虑时间关联性，拟合过于粗糙且计算误差波动较大。多重插补法以贝叶斯理论为基础进行最大期望迭代优化，实现对数据的插补[5-6]。文献[7]利用马尔科夫模型对变压器油中溶解气体数据进行状态空间划分，通过多状态转移补全缺失数据。此外部分学者通过低秩矩阵理论对数据进行分析进而恢复缺失数据[8-9]。上述方法对数据特性有一定的要求，如马尔科夫模型在划分层级时需要考虑数据分布，低秩矩阵理论要求数据具有一定冗余度。

随着新一代机器学习的兴起，部分学者采用人工智能方法补全缺失数据。文献[10]提出了一种深度学习方法解决GPS时间序列数据补全问题；文献[11]提出基于改进生成式对抗网络的数据补全方法；文献[12]通过浅层自动编码器，学习训练网络达到数据补全的目的。但是深度学习等方法在计算过程中需要大量连续、完整的历史数据且需耗费较多的计算资源，更适合连续数据的预测问题。

针对已有的数据补全方法对数据特性有所要求的缺点，本文提出了一种基于灰色自适应*K*-最近邻GAKNN（Grey Adaptive *K*-Nearest Neighbor）方法的数据补全方法：将适量历史数据作为样本库，在朴素*K*-最近邻KNN[13]（*K*-Nearest Neighbor）方法的基础上，构建时间序列特征，通过度量阈值自适应排除相似度不高的特征序列；通过近邻点的灰色关联系数GRC（Grey Relation Coefficient）对近邻点设置权重进而生成更为准确的补全数据。实验结果表明，与其他方法相比，本文方法在样本缺失率为3%、6%、10%及30%时，总体误差均较低。为了验证经过GAKNN方法重建后样本的有效性，将补全的数据样本用于深度学习预测，结果表明利用本文方法得到的预测值更接近真实值，预测误差更低。本文方法无需考虑数据分布情况，仅在适量数据的驱动下即可对配电站时间序列数据的缺失部分进行补全。

1 朴素KNN回归数据补全方法

1.1 时间序列特征构建

配电数据以一定频率依次存储相应时间点的数值，对于频率为$f = 1/t$的时间序列数据，每隔时间t记录1个数据点，则原始数据样本为$\{x(T-nt), x(T-nt+t), \cdots, x(T-t), x(T)\}$（其中，$T$为当前时刻）。通过时间滑动窗口将当前时刻$T$前的$n$个数据点构建为输入序列，然后将该输入序列和当前时刻的数据$x(T)$共同构成训练样本，如式（1）所示。

$$\{x(T-t), x(T-2t), \cdots, x(T-nt), x(T)\} \quad (1)$$

将式（1）记作$\{(\boldsymbol{X}, y)\}$，其中$\boldsymbol{X}=[x(T-t)\ x(T-2t)\ \cdots\ x(T-nt)]$；$y=x(T)$。时间序列的特征

收稿日期：2020-06-03；**修回日期**：2021-06-01
基金项目：国家电网“配电站智能云机器人研究及其应用验证”项目
Project supported by the “Research and Application Verification of Intelligent Cloud Robot in Distribution Station Program” of State Grid Corporation of China

结构如图1所示。

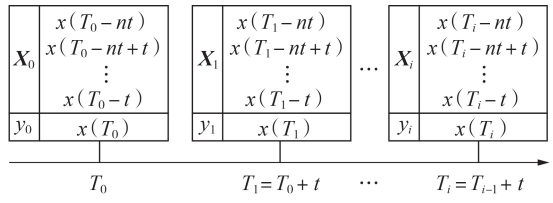


图1 时间序列的特征结构

Fig.1 Feature structure of time series

1.2 基于朴素KNN方法的数据补全过程

朴素KNN方法首先计算输入样本在空间上与训练数据集最近的K个数据点，然后对这K个最近邻数据点取平均值作为输入样本的输出值来补全缺失的数据项。

首先拆分配电站时间序列数据，并按照1.1节构建时间序列特征，获得无缺失项样本和缺失项样本$\boldsymbol{T}_{\text{lack}}$，其中无缺项失样本作为训练数据集$\boldsymbol{T}_{\text{train}}$，如式(2)所示。

$$\boldsymbol{T}_{\text{train}}=\{(\boldsymbol{X}_1,y_1),(\boldsymbol{X}_2,y_2),\cdots,(\boldsymbol{X}_n,y_n)\} \tag{2}$$

式中：$\boldsymbol{X}_i=[x_i^1,x_i^2,\cdots,x_i^m]$，$m$为输入特征的个数，$x_i^j$为第$i$行样本数据中的第$j$个属性；$y_i$为第$i$行样本数据的输出值。

拆分配电站时间序列数据后得到的缺失项样本$\boldsymbol{T}_{\text{lack}}$中仅有输入值而缺失输出值，将其定义为：

$$\boldsymbol{T}_{\text{lack}}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\} \tag{3}$$

通过式(4)计算样本间的距离度量参数d。

$$d(\boldsymbol{X}_i,\boldsymbol{X}_j)=\left[\sum_{l=1}^{m}(x_i^l-x_j^l)^2\right]^{\frac{1}{2}} \tag{4}$$

对$\boldsymbol{T}_{\text{lack}}$中的每个数据进行遍历，在训练集$\boldsymbol{T}_{\text{train}}$中找出与$\boldsymbol{X}_i$最邻近的$K$个数据点，并记作$N_k(x)$，对$N_k(x)$中的$K$个数据点取平均值$\bar{y}$，将$\bar{y}$作为$\boldsymbol{X}_i$的补全输出值，重复上述步骤直至$\boldsymbol{T}_{\text{lack}}$中的数据缺失项全部得到补全。

在朴素KNN方法中，K值的选择会对结果产生重大影响。若K值较小则估计误差会较大，对近邻的实例点较为敏感，整体模型容易发生过拟合；若K值较大则基于较大邻域内的训练数据进行预测，可以减少估计误差，但数据补全易发生错误。另外，朴素KNN方法未对最近邻的K个数据点的权重进行区分，而是使用K个数据点的均值作为预测的结果，所以该方法无法判断各个数据之间的关联程度。且朴素KNN方法计算的$N_k(x)$中的某些点与$\boldsymbol{X}_i(\boldsymbol{X}_i\in\boldsymbol{T}_{\text{lack}})$的距离度量参数大于一定值后，这些点与参考点$\boldsymbol{X}_i$的不相似程度偏高，使用均值计算将增加预测误差，可以考虑设置阈值，自适应地舍去超过阈值的近邻点。

本文通过设置阈值并引入灰色关联系数，舍去$N_k(x)$中与$\boldsymbol{X}_i$的距离超过阈值的近邻点后，计算剩余近邻点的灰色关联系数，根据灰色关联系数计算近邻点的权重，弥补朴素KNN方法中取相同权重的不足。

2 基于GAKNN方法的数据补全

2.1 时间序列数据的灰色关联系数

灰色关联系数根据数据之间发展趋势的相似或者相异程度衡量数据之间的关联程度。本文将其用作除式(4)所示的空间相似度之外的进一步衡量近邻点相似程度的方法。

按照1.1节将原始数据样本构建为式(5)所示的时间序列数据矩阵$\boldsymbol{T}_{\text{f}}$。

$$\boldsymbol{T}_{\text{f}}=\begin{bmatrix}\boldsymbol{X}_1\\\boldsymbol{X}_2\\\vdots\\\boldsymbol{X}_n\end{bmatrix}=\begin{bmatrix}x_1(1)&x_1(2)&\cdots&x_1(m)\\x_2(1)&x_2(2)&\cdots&x_2(m)\\\vdots&\vdots&&\vdots\\x_n(1)&x_n(2)&\cdots&x_n(m)\end{bmatrix} \tag{5}$$

式中：$x_i(k)=x(T_i-(m-k+1)t)$，即当前时刻T_i向前推移$(m-k+1)t$的数据值。

根据朴素KNN思想，将缺失数据的前m个数据序列作为参考数据$\boldsymbol{X}_0$，如式(6)所示。

$$\boldsymbol{X}_0=[x_0(1)\ \ x_0(2)\ \ \cdots\ \ x_0(m)] \tag{6}$$

按照式(7)对时间序列数据进行无量纲化处理，形成式(8)所示的无量纲矩阵$\boldsymbol{T}'_{\text{f}}$。

$$x'_i(k)=\frac{x_i(k)}{\dfrac{1}{m}\displaystyle\sum_{k=1}^{m}x_i(k)} \tag{7}$$

$$\boldsymbol{T}'_{\text{f}}=\begin{bmatrix}\boldsymbol{X}'_1\\\boldsymbol{X}'_2\\\vdots\\\boldsymbol{X}'_n\end{bmatrix}=\begin{bmatrix}x'_1(1)&x'_1(2)&\cdots&x'_1(m)\\x'_2(1)&x'_2(2)&\cdots&x'_2(m)\\\vdots&\vdots&&\vdots\\x'_n(1)&x'_n(2)&\cdots&x'_n(m)\end{bmatrix} \tag{8}$$

$$\boldsymbol{X}'_0=\begin{bmatrix}x'_0(1)&x'_0(2)&\cdots&x'_0(m)\end{bmatrix} \tag{9}$$

按照式(10)计算$\boldsymbol{X}'_i(\boldsymbol{X}'_i\in\boldsymbol{T}'_{\text{f}})$与参考数据$\boldsymbol{X}'_0$在第$l$个输入特征属性的灰色关联系数$\lambda_{\text{GRC}}(x'_0(l),x'_i(l))$。

$$\lambda_{\text{GRC}}(x'_0(l),x'_i(l))=\frac{\min\limits_{\forall i}\min\limits_{\forall k}\left|x'_0(k)-x'_i(k)\right|+\rho\max\limits_{\forall i}\max\limits_{\forall k}\left|x'_0(k)-x'_i(k)\right|}{\left|x'_0(l)-x'_i(l)\right|+\rho\max\limits_{\forall i}\max\limits_{\forall k}\left|x'_0(k)-x'_i(k)\right|} \tag{10}$$

式中：$0<\rho<1$，为分辨系数，表示灰色关联系数间的差异程度，通常取为0.5。

针对所有输入特征属性，按照式(11)计算$\boldsymbol{X}'_i$与参考数据$\boldsymbol{X}'_0$的灰色关联系数的平均值$\lambda_{\text{GRC}}(\boldsymbol{X}'_0,\boldsymbol{X}'_i)$，该值即表示两者之间的关联程度。

$$\lambda_{\text{GRC}}(\boldsymbol{X}'_0,\boldsymbol{X}'_i)=\frac{1}{m}\sum_{l=1}^{m}\lambda_{\text{GRC}}(x'_0(l),x'_i(l)) \tag{11}$$

2.2 基于GAKNN方法的数据补全流程

基于GAKNN方法的数据补全流程如附录A图

A1所示，图中浅色方框内为该方法的作用范围。本节对流程进行简要说明，详细步骤见2.3节。

1）根据时间序列数据是否完整将原始数据划分为连续数据段和缺失数据段，并通过时间滑动窗口构建时间序列数据$\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$，其中连续数据段$\boldsymbol{T}_{\text{train}}$作为训练数据，缺失数据段$\boldsymbol{X}_{\text{lack_}i}$作为第$i$个缺失项，为需要补全的部分。

2）基于2.1节中的朴素KNN方法，在训练集中依次检索，得到初步缺失数据$\boldsymbol{X}_s(\boldsymbol{X}_s \in \boldsymbol{X}_{\text{lack_}i})$的$K$个近邻点，然后删除其中超出阈值的数据点，将剩余的近邻点记为$\boldsymbol{T}_1=\{(\boldsymbol{X}_0, y_0),(\boldsymbol{X}_1, y_1), \cdots,(\boldsymbol{X}_M, y_M)\}$，其中$M$为剩余近邻点个数。

3）计算剩余近邻点的灰色关联系数$\lambda_{\text{GRC}}(\boldsymbol{X}_s, \boldsymbol{X}_i)$，并计算各近邻点的权重$w_i$，最后通过对各近邻点进行线性组合得出缺失项的数据$\left\{y=\sum w_i y_i\right\}$。

对于缺失数据段依次进行上述步骤直至补全所缺失的全部数据。

2.3　基于GAKNN方法的数据补全步骤

1）原始配电数据样本的预处理。将配电站时间序列数据样本拆分为缺失项和无缺失项，并构建时间序列特征，得到缺失项$\boldsymbol{T}_{\text{lack}}$以及训练项$\boldsymbol{T}_{\text{train}}$，其中缺失项$\boldsymbol{T}_{\text{lack}}$定义为：

$$\boldsymbol{T}_{\text{lack}}=\{\boldsymbol{X}_{\text{lack\_}1}, \boldsymbol{X}_{\text{lack\_}2}, \cdots, \boldsymbol{X}_{\text{lack\_}r}\} \tag{12}$$

式中：r为缺失项的总数。

对于任意$\boldsymbol{X}_{\text{lack_}i} \in \boldsymbol{T}_{\text{lack}}$应满足：

$$\boldsymbol{X}_{\text{lack\_}i}=\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\\ \vdots\\ \boldsymbol{X}_p\end{bmatrix}=\begin{bmatrix}x_1(1) & x_1(2) & \cdots & x_1(m)\\ x_2(1) & x_2(2) & \cdots & x_2(m)\\ \vdots & \vdots & & \vdots\\ x_p(1) & x_p(2) & \cdots & x_p(m)\end{bmatrix} \tag{13}$$

式中：p为缺失项样本数。

训练项样本$\boldsymbol{T}_{\text{train}}=\{(\boldsymbol{X}_{\text{train}}, \boldsymbol{Y}_{\text{train}})\}$，$\boldsymbol{X}_{\text{train}}$、$\boldsymbol{Y}_{\text{train}}$分别如式(14)、(15)所示。

$$\boldsymbol{X}_{\text{train}}=\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\\ \vdots\\ \boldsymbol{X}_n\end{bmatrix}=\begin{bmatrix}x_1(1) & x_1(2) & \cdots & x_1(m)\\ x_2(1) & x_2(2) & \cdots & x_2(m)\\ \vdots & \vdots & & \vdots\\ x_n(1) & x_n(2) & \cdots & x_n(m)\end{bmatrix} \tag{14}$$

$$\boldsymbol{Y}_{\text{train}}=[\,y_1 \;\; y_2 \;\; \cdots \;\; y_n\,]^{\text{T}} \tag{15}$$

2）在训练数据中，获得与缺失项最近邻的K个数据。首先按照式(16)计算缺失项$\boldsymbol{X}_s(\boldsymbol{X}_s \in \boldsymbol{X}_{\text{lack_}i})$与$\forall \boldsymbol{X}_u(\boldsymbol{X}_u \in \boldsymbol{X}_{\text{train}})$的欧氏距离。

$$d(\boldsymbol{X}_s, \boldsymbol{X}_u)=\sqrt{\sum_{k=1}^{m}(x_s(k)-x_u(k))^2} \tag{16}$$

$$\boldsymbol{X}_s=[\,x_s(1) \;\; x_s(2) \;\; \cdots \;\; x_s(m)\,] \quad \boldsymbol{X}_s \in \boldsymbol{X}_{\text{lack\_}i}$$

将训练数据中与缺失项$\boldsymbol{X}_s$最近邻的K个数据构成矩阵，如式(17)所示。

$$\boldsymbol{X}_{\text{close}}=\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\\ \vdots\\ \boldsymbol{X}_K\end{bmatrix}=\begin{bmatrix}x_1(1) & x_1(2) & \cdots & x_1(m)\\ x_2(1) & x_2(2) & \cdots & x_2(m)\\ \vdots & \vdots & & \vdots\\ x_K(1) & x_K(2) & \cdots & x_K(m)\end{bmatrix} \tag{17}$$

$$\boldsymbol{Y}_{\text{close}}=[\,y_1 \;\; y_2 \;\; \cdots \;\; y_n\,]^{\text{T}} \tag{18}$$

3）舍去最近邻的K个数据中不相似度偏高的数据点。计算$d_{si}=d(\boldsymbol{X}_s, \boldsymbol{X}_i)(\boldsymbol{X}_i \in \boldsymbol{X}_{\text{close}})$，若满足$d_{si} \leqslant D$（$D$为阈值），则保留$\boldsymbol{X}_i$，否则舍去$\boldsymbol{X}_i$。经过自适应操作后，最近邻数据的数量变为$M$，从而得到矩阵如式(19)、(20)所示。

$$\boldsymbol{X}'_{\text{close}}=\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\\ \vdots\\ \boldsymbol{X}_M\end{bmatrix}=\begin{bmatrix}x_1(1) & x_1(2) & \cdots & x_1(m)\\ x_2(1) & x_2(2) & \cdots & x_2(m)\\ \vdots & \vdots & & \vdots\\ x_M(1) & x_M(2) & \cdots & x_M(m)\end{bmatrix} \tag{19}$$

$$\boldsymbol{Y}'_{\text{close}}=[\,y_1 \;\; y_2 \;\; \cdots \;\; y_M\,]^{\text{T}} \tag{20}$$

4）将缺失项$\boldsymbol{X}_s$作为参考项，根据式(21)计算$\boldsymbol{X}_j(\boldsymbol{X}_j \in \boldsymbol{X}'_{\text{close}})$与$\boldsymbol{X}_s$的灰色关联系数。

$$\lambda_{\text{GRC}}(\boldsymbol{X}_s, \boldsymbol{X}_j)=\frac{1}{m}\sum_{l=1}^{m}\lambda_{\text{GRC}}(x_s(l), x_j(l)) \tag{21}$$

根据灰色关联系数，按式(22)计算M个最近邻数据的权重系数。

$$w_k=\frac{\lambda_{\text{GRC}}(\boldsymbol{X}_s, \boldsymbol{X}_k)}{\sum_{j=1}^{M}\lambda_{\text{GRC}}(\boldsymbol{X}_s, \boldsymbol{X}_j)} \tag{22}$$

5）按照式(23)将式(18)所得数据与式(22)计算得到的权重进行线性组合，得到填补值y。

$$y=\sum_{i=1}^{K}w_i y_i \tag{23}$$

将填补值y补充到$\boldsymbol{X}_s$中。若$\boldsymbol{X}_{\text{lack_}i}$未补充完全，即$s<p$，则对于$\boldsymbol{X}_{s+1}$重复步骤2)—5)；若$\boldsymbol{X}_{\text{lack_}i}$被补充完全，即$s=p$，则对于$\boldsymbol{X}_{\text{lack_}i+1}$重复步骤2)—5)直至$\boldsymbol{T}_{\text{lack}}$被补充完全。

3　实验结果分析

3.1　实验数据

实验数据采用江苏省某市配电房的某配电柜在2019年9月份的现场运行数据。数据记录的频率为每分钟记录1个数据点，数据点以电流（单位为A）的形式保存。数据样本有30 120条，其中，前25 000条为完整数据，可作为训练样本；对后5 120条数据进行随机缺失处理，使得数据缺失率分别达到3%、6%、10%和30%，作为待补全的对象。

3.2　参数选择

为了获取基于GAKNN方法的数据补全方法最合适的K值，利用缺失率不同的样本研究K值对数据补全结果的影响，采用均方根误差RMSE（Root

Mean Squared Error)以及平均绝对百分误差MAPE(Mean Absolute Percentage Error)作为数据补全结果的评价指标,分别如式(24)、(25)所示。

$$E_{\mathrm{RMSE}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2} \tag{24}$$

$$E_{\mathrm{MAPE}}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \tag{25}$$

式中:E_{RMSE}和E_{MAPE}分别为RMSE和MAPE的值;$\hat{y}_i$为预测值;y_i为样本中的真实值。

在不同缺失率下,K值对数据补全结果的影响如图2所示。由图可见:不同的缺失率下的数据补全结果有一定的差异;整体而言在K值较小时RMSE可能出现波动,但随着K值到达一定值后RMSE逐渐增大,最后趋于稳定。当K=7时,在不同的缺失率下均有较好的数据补全结果。因此,本文取K=7。

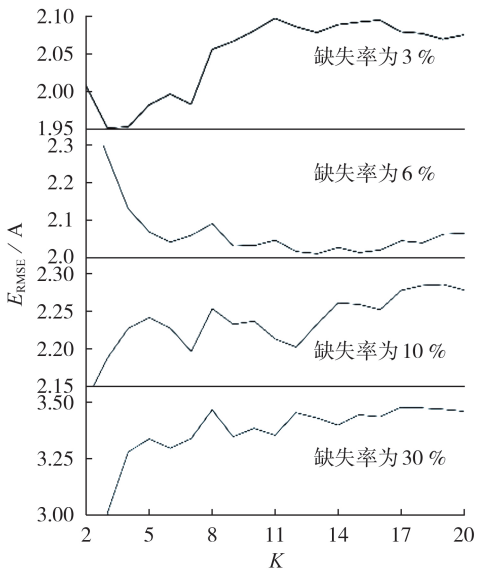


图2 不同缺失率下K值对数据补全效果的影响

Fig.2 Influence of K on data completion results under different missing rates

当K=7时,不同缺失率下的数据补全结果与真实值的对比如图3所示。图中,t_1—t_5分别对应9月12日19:40、9月12日23:00、9月13日02:20、9月13日05:40、9月13日09:00。当缺失率在3%~30%范围内时,补全后的数据在大部分情况下能覆盖原数据,且两者间的误差控制在8%以下;但当样本缺失率过高后,数据补全的效果变差,则此样本不适合继续用于数据补全,可考虑舍去或另外进行采样。

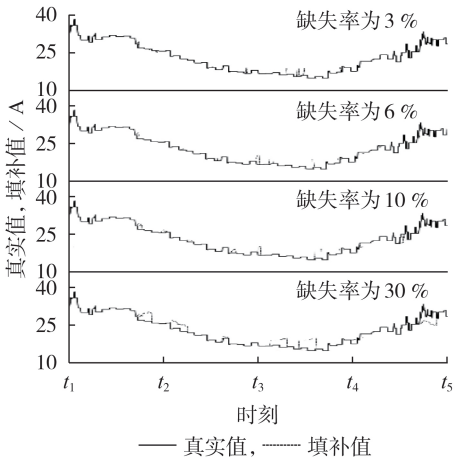


图3 不同缺失率下的数据补全结果与真实值的对比

Fig.3 Comparison between data completion results and real values under different missing rates

3.3 数据补全结果对比实验分析

参数确定后,比较本文方法和均值、马尔科夫、朴素KNN以及ARIMA方法的数据补全结果与真实值的误差,实验评价指标采用RMSE和MAPE,对比实验结果如表1所示。由表可见,本文方法的RMSE和MAPE的值相对较小。其中在样本缺失率为3%时,本文方法的RMSE值为1.997 A,仅比ARIMA方法的RMSE值高;当样本缺失率为6%、10%及30%时,本文方法的RMSE值分别为2.024、2.228、3.295 A,与其他4种方法相比均处于较低水平。

随着电力大数据挖掘成为热点,电力负荷的预测受到学者的关注,数据补全的一个关键作用是将补全后的完整样本用于后续的深度学习数据挖掘。设置缺失率为3%,将采用上述5种方法进行数据补全后的样本用于长短期记忆LSTM(Long Short-Term Memory)网络[14-15]预测,LSTM网络结构具体参数如表2所示,LSTM网络的预测结果如附录A图A2所示。

图4给出了不同缺失率下,LSTM网络预测结果的RMSE和MAPE值。由图4可见,本文方法的RMSE、MAPE值整体相对较小,即填补的样本更接近真实样本,对于深度学习数据挖掘的误差更小,预测的值更趋于真实结果,进一步验证了本文方法的有效性。

表1 不同方法的数据补全结果对比

Table 1 Comparison of data completion results under different methods

缺失率/%	均值方法的结果		马尔科夫方法的结果		朴素KNN方法的结果		ARIMA方法的结果		本文方法的结果	
	E_{RMSE} / A	E_{MAPE}	E_{RMSE} / A	E_{MAPE}	E_{RMSE} / A	E_{MAPE}	E_{RMSE} / A	E_{MAPE}	E_{RMSE} / A	E_{MAPE}
3	45.946	2.077	2.808	0.162	2.082	0.098	0.918	0.026	1.997	0.059
6	88.608	3.809	3.517	0.204	2.021	0.090	2.805	0.224	2.024	0.053
10	148.633	5.947	4.888	0.251	2.209	0.090	4.663	0.472	2.228	0.066
30	479.800	17.632	3.855	0.195	3.404	0.136	14.700	2.607	3.295	0.087

表2　LSTM网络参数

Table 2　Parameters of LSTM network

网络结构	超参数
输入序列长度	20
输出序列长度	1
LSTM网络层数	3
LSTM网络隐藏节点	[64]
迭代次数	100
损失函数	MSE
训练优化器	RMSProp

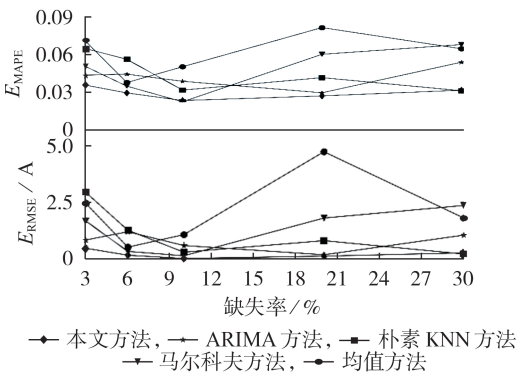


图4　LSTM网络的预测误差

Fig.4　Prediction error of LSTM network

4　结论

本文针对配电站时间序列的数据缺失问题，提出了一种基于GAKNN方法的数据补全方法。该方法在朴素KNN方法的基础上，拆分构建的时间序列特征，设置距离度量阈值实现近邻点K值的自适应选择，然后利用灰色关联系数计算近邻点的权重，最后通过线性组合补全缺失数据。在实验中结合江苏省某市的电力数据进行算例分析，讨论了参数选择对数据补全误差的影响，并将本文方法与其他4种方法进行对比，结果表明本文方法的补全数据误差更低。此外将多种方法进行数据补全后的样本用于深度学习预测，结果表明利用本文方法进行数据补全后的样本在深度学习预测方面表现更佳。

本文探讨的缺失数据仅限正常数据的记录缺失，而未考虑配电设备异常、引发切断而造成的缺失。本文方法适用于数据缺失率低于30%、出现数据断层且需要补全数据用于进一步数据挖掘的场景。后续工作将围绕数据补全后的深度模型在线更新、优化等方面展开，探索数据补全与深度学习预测方面的关联。

附录见本刊网络版(http://www.epae.cn)。

参考文献：

[1] 孙国强，沈培锋，赵扬，等. 融合知识库和深度学习的电网监控告警事件智能识别[J]. 电力自动化设备，2020，40(4)：40-47.
SUN Guoqiang，SHEN Peifeng，ZHAO Yang，et al. Intelligent recognition of power grid monitoring alarm event combining knowledge base and deep learning[J]. Electric Power Automation Equipment，2020，40(4)：40-47.

[2] 程万伟. 时间序列缺失值插补方法研究[D]. 长沙：湖南大学，2018.
CHENG Wanwei. Study on the interpolation method of time series missing value[D]. Changsha：Hunan University，2018.

[3] SUN Jian，LIAO Haitao，UPADHYAYA B R. A robust functional-data-analysis method for data recovery in multi-channel sensor systems[J]. IEEE Transactions on Cybernetics，2014，44(8)：1420-1431.

[4] FARHANGFAR A，KURGAN L，DY J. Impact of imputation of missing values on classification error for discrete data[J]. Pattern Recognition，2008，41(12)：3692-3705.

[5] RUBIN D B. Multiple imputation after 18+ years[J]. Journal of the American Statistical Association，1996，91(434)：473-489.

[6] CAMPION W M，RUBIN D B. Multiple imputation for nonresponse in surveys[J]. Journal of Marketing Research，1989，26(4)：485.

[7] 张若愚，齐波，张鹏，等. 面向电力变压器状态评价的油中溶解气体监测数据补全方法[J]. 电力自动化设备，2019，39(11)：181-187.
ZHANG Ruoyu，QI Bo，ZHANG Peng，et al. Method for interpolating monitoring data of dissolved gas in oil for power transformer state assessment[J]. Electric Power Automation Equipment，2019，39(11)：181-187.

[8] ASHRAPHIJUO M，AGGARWAL V，WANG X D. A characterization of sampling patterns for low-tucker-rank tensor completion problem[C]//2017 IEEE International Symposium on Information Theory(ISIT). Aachen，Germany：IEEE，2017：531-535.

[9] 乔文俞，李野，刘浩宇，等. 基于曲线相似与低秩矩阵的缺失电量数据补全方法[J]. 电力建设，2020，41(1)：32-38.
QIAO Wenyu，LI Ye，LIU Haoyu，et al. Missing load data completion based on curve similarity and low-rank matrix[J]. Electric Power Construction，2020，41(1)：32-38.

[10] 尹玲，尹京苑，孙宪坤，等. 缺失GPS时间序列的神经网络补全[J]. 测绘科学技术学报，2018，35(4)：331-336.
YIN Ling，YIN Jingyuan，SUN Xiankun，et al. Reconstruction of gappy GPS coordinate time series based on long short-term memory networks[J]. Journal of Geomatics Science and Technology，2018，35(4)：331-336.

[11] 王守相，陈海文，潘志新，等. 采用改进生成式对抗网络的电力系统量测缺失数据重建方法[J]. 中国电机工程学报，2019，39(1)：56-64.
WANG Shouxiang，CHEN Haiwen，PAN Zhixin，et al. A reconstruction method for missing data in power system measurement using an improved generative adversarial network[J]. Proceedings of the CSEE，2019，39(1)：56-64.

[12] MIRANDA V，KRSTULOVIC J，KEKO H，et al. Reconstructing missing data in state estimation with autoencoders[J]. IEEE Transactions on Power Systems，2012，27(2)：604-611.

[13] 张孙力，杨慧中. 基于改进的K近缺失数据补全[J]. 计算机与应用化学，2015，32(12)：1499-1502.
ZHANG Sunli，YANG Huizhong. Missing data completion based on an improved K-neighbor algorithm[J]. Computers and Applied Chemistry，2015，32(12)：1499-1502.

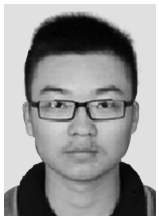
[14] HOCHREITER S，SCHMIDHUBER J. Long short-term memory[J]. Neural Computation，1997，9(8)：1735-1780.

[15] 刘俐，李勇，曹一家，等. 基于支持向量机和长短期记忆网络的暂态功角稳定预测方法[J]. 电力自动化设备，2020，40(2)：

129-139.
LIU Li, LI Yong, CAO Yijia, et al. Transient rotor angle stability prediction method based on SVM and LSTM network[J]. Electric Power Automation Equipment, 2020, 40(2): 129-139.

作者简介:

冯 磊(1995—),男,江西抚州人,硕士研究生,研究方向为电力负荷预测、深度学习(**E-mail**:lei_lei@sjtu.edu.cn);



冯 磊

王石刚(1957—),男,湖北黄石人,教授,博士研究生导师,研究方向为电力机器人、复杂机电系统设计与研发(**E-mail**:wangshigang@sjtu.edu.cn);

梁庆华(1972—),男,江苏姜堰人,副教授,研究方向为机电一体化设计(**E-mail**:qhliang@sjtu.edu.cn)。

(**编辑** 任思思)

Completion method for missing time series data of distribution station based on GAKNN method

FENG Lei, WANG Shigang, LIANG Qinghua

(School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Data record loss may occur in the process of time series data acquisition, transmission and storage in distribution station, which affects high-level data analysis and processing to a certain extent. To solve this problem, a completion method for missing data based on GAKNN (Gray Adaptive *K*-Nearest Neighbor) method is proposed. Firstly, the time series features are constructed. Then, the nearest neighbor points are selected by the threshold based on simple KNN (*K*-Nearest Neighbor) method, and the weight coefficients of the nearest neighbor points are calculated combining with the gray correlation coefficient. Finally, the missing data can be completed in turn. With the electric power data sample of a city in Jiangsu Province, the test results show that the missing data completion results of GAKNN method are better than those of other methods, and the completed samples have lower errors in deep learning prediction.

Key words: distribution station data; time series; data completion; GAKNN

(上接第 177 页 continued from page 177)

Fault diagnosis method of rolling bearing based on MVMD and full vector envelope spectrum

HUANG Chuanjin[1], SONG Haijun[1], YANG Shixi[2], CHI Yongwei[2], HUANG Haizhou[3], HAO Shuang[1], GUO Shengbin[1]

(1. School of Mechanical and Electrical Engineering, Zhengzhou Institute of Technology, Zhengzhou 450044, China;
2. School of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China;
3. Huadian Electric Power Research Institute Co., Ltd., Hangzhou 310030, China)

Abstract: In order to diagnose the rolling bearing faults comprehensively and accurately, a fault diagnosis method of rolling bearing based on MVMD (Multivariate Variational Mode Decomposition) and full vector envelope spectrum is proposed. Firstly, the quadrature sampling technology is used to obtain the vibration signals in the mutually perpendicular directions at the same support of the rolling bearing, which are formed into a binary modulation oscillation signal. Then, MVMD is used to extract a set of optimal binary modulation oscillation signals from the binary modulation oscillation signal, the sum of the corresponding bandwidth is the smallest. Because MVMD uses a unified mathematical model to build the signal models in two directions, it can ensure that the fault features are decomposed to the same layer to facilitate subsequent information fusion. Finally, the Hilbert transformation is used to demodulate each binary modulation oscillation signal to obtain the corresponding envelope signal. The envelope signal information in two directions is fused by full vector spectrum to obtain the full vector envelope spectrum to diagnose the rolling bearing fault. Simulative and test results prove the feasibility and effectiveness of the proposed method.

Key words: multivariate variational mode decomposition; rolling bearing; fault diagnosis; full vector envelope spectrum

附录 A

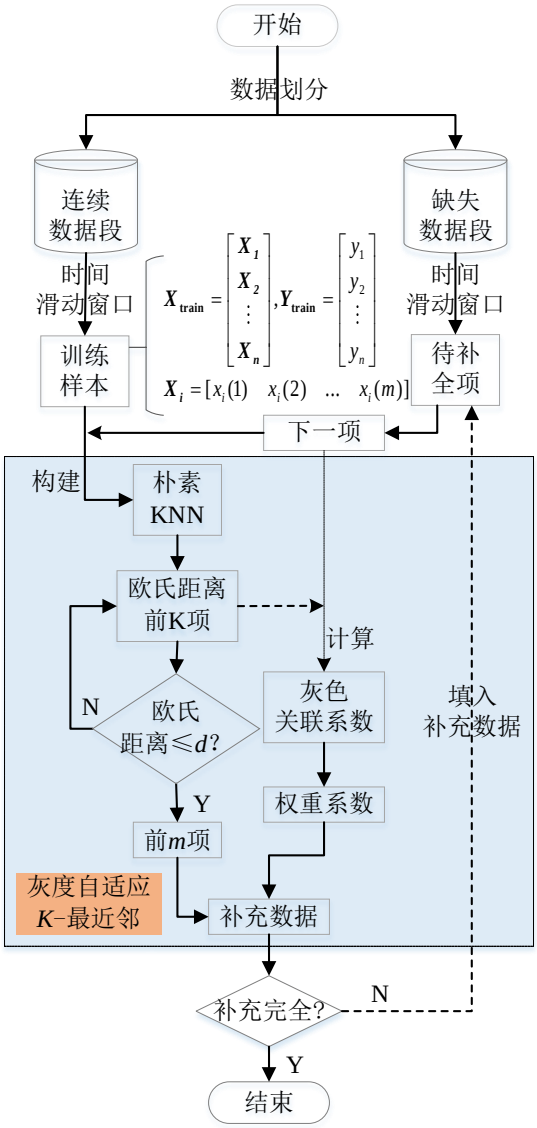


图 A1 基于 GAKNN 的数据补全方法流程

Fig.A1 Flowchart of data completion method based on GAKNN

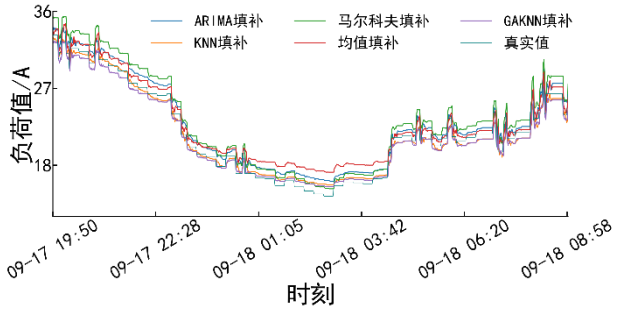


图 A2 补全样本的 LSTM 预测误差

Fig.A2 LSTM prediction error under complete samples