计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191200152

基于稀疏表示的电力负荷数据补全

李培冠[1] 於志勇[1,2] 黄昉菀[1,2]
1 福州大学数学与计算机科学学院 福州 350116
2 福州大学福建省网络计算与智能信息处理重点实验室 福州 350116
(lipeiguan163@163.com)

摘 要 数据缺失在电力负荷数据采集过程中经常发生,对提高算法的预测精确度带来了不利影响。现有的缺失数据补全算法只适用于缺失数据量较少的情况,而对于缺失数据较多的情况表现不佳。面对严重数据缺失的挑战,文中提出了一种基于稀疏表示的电力负荷缺失数据补全方法。首先以数据随机缺失为前提,将训练数据中假定缺失后的数据与完整的训练数据上下拼接构成训练矩阵;其次,利用离散余弦变换(Discrete Cosine Transform,DCT)生成一个过完备字典,并根据训练矩阵对其进行学习,旨在通过调优得到一个合适的字典,能对训练矩阵中的样本进行最好的稀疏表示。最后,在测试阶段,先利用学习后字典的上半部分获得测试集缺失数据的稀疏表示,然后利用稀疏表示和学习后字典的下半部分重构出无缺失的完整数据。实验结果表明,使用该方法对电力负荷数据缺失值进行补全,可以获得比传统插值方法、基于相关性的KNN算法、时空压缩感知估计算法以及时序压缩感知预测算法更高的精度。即使数据缺失率高达95%,该方法依然可以有效地补全缺失数据。

关键词:电力负荷;缺失数据;稀疏表示;数据补全
中图法分类号 TP181

Power Load Data Completion Based on Sparse Representation

LI Pei-guan[1],YU Zhi-yong[1,2] and HUANG Fang-wan[1,2]
1 College of Mathematics and Computer Science,Fuzhou University,Fuzhou 350116,China
2 Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing,Fuzhou University,Fuzhou 350116,China

Abstract Data loss often occurs in the process of power load data collection,which adversely affects the accuracy of algorithm prediction. The existing missing data completion algorithm is only suitable for the case with less missing data,but performs poorly for the case with more missing data. Faced with the challenge of severe data loss,a method for power load missing data completion based on sparse representation is proposed. First of all,we assume that the data is randomly missing,and stitch the assumed missing data in the training data and the complete training data to form a training matrix. Secondly,an over-complete dictionary is generated by discrete cosine transform (DCT),and is learned according to the training matrix,aims to obtain a suitable dictionary for the best sparse representations of the samples in the training matrix. Finally,in the test phase,the upper part of the learned dictionary is used to obtain sparse representations of the missing data in the test set,and then the sparse representations and the lower part of the learned dictionary are used to reconstruct the complete data without missing. Experimental results show that using this method to complete missing values of power load data can achieve higher accuracy than traditional interpolation methods,correlation-based KNN algorithm,spatiotemporal compressed sensing estimation algorithm and time-series compressed sensing prediction algorithm. Even if the data miss rate is as high as 95%,this method can still effectively complete the missing data.

Keywords Power load,Missing data,Sparse representation,Data completion

1 引言

电力负荷预测是供电部门的重要工作之一。准确的负荷预测可以经济合理地安排电网内部发电机组的启停,保持电网运行的安全稳定性,减少不必要的旋转储备容量,合理安排机组检修计划,保证社会的正常生产和生活,有效地降低发电成本并提高经济效益和社会效益[1-4]。智能电表等智能采集设备在智能电网中的不断普及为精准预测电力负荷提供了很大的帮助。然而,由于设备损坏、外界气候等原因造成的数据缺失情况在工业数据集中特别严重,如文献[5]的数据集中几

到稿日期:2019-12-25 返修日期:2020-04-16 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61772136);福建省中青年教师教育科研项目(JT180045)
This work was supported by the National Natural Science Foundation of China(61772136) and Research Project for Young and Middle-aged Teachers of Fujian Province (JT180045).
通信作者:黄昉菀(hfw@fzu.edu.cn)

十个属性缺失率均超过 50%，几乎每一条记录都含有缺失值。因此在电力负荷数据的分析与利用中，为了能够更加充分地利用已经搜集到的数据，对缺失数据进行处理是非常必要的。但是，现有方法只适用于数据集中具有少量缺失值的情况，当数据缺失率很高时往往表现不佳。为了解决上述问题，本文提出了一种基于稀疏表示的电力负荷缺失补全方法（Sparse Representation Completion Method，SRCM）来有效恢复存在严重缺失的电力负荷数据。

2 相关工作

目前，已经有大量的方法致力于缺失数据的补全[6]。基于插值的方法是最简单的缺失数据补全算法，如线性插值[7]、三次样条插值[8-9]和 Hermite 插值[10]等。上述方法因容易实现、计算简洁，常被用于补全缺失数据，但随着缺失数据的增多，其效果往往无法令人满意。

近年来，基于机器学习的方法得到了越来越多的关注，成为补全缺失数据最常用的方法。文献[11]使用 KNN 补全算法来补全缺失值，尽管 KNN 补全算法具有简单直观、易于实现、无需先验知识等特点，但其补全的精度取决于近邻样本数据的平均值，存在一定不足之处。文献[12]使用经典的自回归积分滑动平均模型（Auto Regressive Integrated Moving Average Model，ARIMA）来补全缺失数据，虽然取得了不错的效果，但是其并没有充分利用数据缺失之后采集的数据，从而影响了补全效果。文献[13]利用支持向量机（Support Vector Machines，SVM）构造缺失数据补全框架，并有效地补全了电网监测数据中的缺失值。但是，SVM 并不适用于缺失数据较多的情况。压缩感知（Compressed Sensing，CS）算法则是利用信号本身或在某个变换域下的稀疏性，实现对信号的欠采样，从而利用一小部分数据恢复整个数据集[14]。对于缺失值补全问题，文献[15]通过基于主成分分析（Principle Component Analysis，PCA）的实际探测数据的实证研究，观察到道路网络的交通状况中存在隐藏的结构，由此提出了一种时空压缩感知估计算法来解决缺失的数据问题。该算法利用隐藏的结构来计算道路交通状况的估计值。文献[16]提出了一种基于压缩感知的缺失数据预测算法。该算法首先利用时间序列的时域平滑特性设计稀疏表示基，从而将缺失数据预测问题转化成稀疏向量恢复问题，并根据未缺失数据的位置特点设计了与稀疏表示基相关性低的观测矩阵，从而保证了算法的重构性能。文献[17]利用基于压缩感知的方法实现了网络流量估计。虽然基于压缩感知的方法在以上数据集中有着比较不错的效果，但是在电力负荷数据大量缺失的情况下，依然表现的不尽人意，这与 CS 方法要求数据集具有固有的结构有关。文献[18]提出了一种通用方法，该方法采用非负矩阵分解方法将列和行的辅助信息包含在常规线性测量中，并提出了一种新算法（Hierarchical Alternating Least Squares with eXogeneous variables，HALSX）。文献[19]提出了一种用于恢复心音信号的方法。文献[20]探讨了基于递归神经网络（Recurrent Neural Network，RNN）处理缺失数据的方法。然而，RNN 需要大量的训练数据，因此当数据缺失率较高时，RNN 很难发现数据的潜在规律。

通过上述分析，不难看出目前国内外对数据补全技术的研究存在一些普遍性的问题：1）上述方法主要是针对小规模数据集，对于大规模数据集，存在计算效率较低的问题；2）现有方法对于仅具有少量缺失值的数据集表现较好，当数据集的缺失率很高时往往表现不佳。在过去几年中，稀疏表示越来越受到研究人员的关注，尤其是在信号处理、图像处理、机器学习等方面显示出巨大的能力[21]。为了解决上述问题，本文提出了一种基于稀疏表示的电力负荷缺失补全方法 SRCM 来解决存在严重数据缺失的电力负荷补全问题。

3 基于稀疏表示的数据补全方法

基于稀疏表示的数据补全方法的特点在于其并不依赖于初始过完备字典的性能，而依赖于训练矩阵对字典学习的效果，因此对于不同数据集或不同缺失情况均具有良好的适应性。下面简要介绍稀疏表示原理以及 SRCM 的具体步骤。

3.1 稀疏表示简介

给定一条时间序列 $\boldsymbol{y}\in R^m$，构建一个字典 $\boldsymbol{D}\in R^{m\times k}$，若字典的列数 k 大于行数 m，则将该字典称为过完备字典。通常将字典 $\boldsymbol{D}$ 中的一列称为基向量或者“原子”。稀疏表示的主要思想是通过构建合适的过完备字典，利用较少原子的线性组合来表示大部分或全部原始信号。求解时间序列 $\boldsymbol{y}$ 的稀疏表示 $\boldsymbol{\alpha}\in R^m$ 的目标函数一般定义为：

$$\min_{\alpha}\|\boldsymbol{\alpha}\|_0 \quad \text{s.t. } y=D\alpha \tag{1}$$

其中，$\|\cdot\|_0$ 表示 $\boldsymbol{\alpha}$ 中非零的个数，其值越小，表示 $\boldsymbol{\alpha}$ 越稀疏。考虑到真实数据往往存在一定的噪声，目标函数也可以用式(2)表示：

$$\min_{\alpha}\|\boldsymbol{\alpha}\|_0 \text{ s.t. } \|y-D\alpha\|_2^2\leqslant\delta \tag{2}$$

其中，$\boldsymbol{\delta}$ 可以认为是重构误差或噪声信号。通过翻转目标函数和约束条件，式(2)可以转换为：

$$min_{\alpha}\|y-D\alpha\|_2^2\leqslant\delta \quad s.t. \ \|\boldsymbol{\alpha}\|_0\leqslant l \tag{3}$$

其中，稀疏度 $\boldsymbol{l}$ 表示向量 $\boldsymbol{\alpha}$ 中非零项的最大个数。

3.2 SRCM 算法描述

定义 1 完整数据矩阵一个 $m\times n$ 的矩阵，表示一个完整的电力负荷数据集。矩阵中每个数据点都是有效的，即没有缺失的数据。

定义 2 部分数据矩阵是一个 $m\times n$ 的矩阵，记录真实收集的电力负荷数据。其缺失数据置为 0，具体的缺失位置由定义 3 给出。

定义 3 采集位置矩阵（$\boldsymbol{S}$）是一个 $m\times n$ 的矩阵，表示部分数据矩阵中的数据点是否缺失。若 $\boldsymbol{s}(i,j)=0$，则意味着部分数据矩阵中该位置的数据缺失；若 $\boldsymbol{s}(i,j)=1$，则意味着部分数据矩阵中该位置的数据未出现缺失。

定义 4 补全的完整矩阵：通过在部分数据矩阵中补全缺失值生成的补全的完整矩阵。

SRCM 的结构框图如图 1 所示，其图形化的实现过程如图 2 所示，共包含 5 个步骤：数据预处理、字典构造、字典学习、稀疏表示和数据重构。其中，完整数据矩阵表示为 $\boldsymbol{X}_{\text{full}}$，$\boldsymbol{X}_{\text{loss}}$ 与 $\boldsymbol{Y}_{\text{loss}}$ 表示部分数据矩阵，$\boldsymbol{Y}'_{\text{full}}$ 表示补全的完整矩阵。在图 2 中，灰色方块表示有数据，白色方块表示数值 0。

步骤 1 首先将部分数据矩阵 $\boldsymbol{X}_{\text{loss}}$ 与完整数据矩阵 X_{full}

采用上下拼接的方式构建出训练矩阵 $\boldsymbol{X}\in\boldsymbol{R}^{2m\times n}$。

步骤 2 根据 $\boldsymbol{X}$ 的行数，生成与之对应的过完备字典 $\boldsymbol{D}\in R^{2m\times k}$（$k\gg 2m$）。可以把 $\boldsymbol{D}$ 的上半部分称为 $\boldsymbol{D}_1\in R^{m\times k}$，把下半部分称为 $\boldsymbol{D}_2\in R^{m\times k}$。

步骤 3 利用 K-SVD 算法对字典进行学习，得到优化后的字典 $\boldsymbol{D}'$。字典学习的目的在于学习一个对于训练集而言具有最小重构误差的字典。

步骤 4 给出部分数据矩阵 $\boldsymbol{Y}_{\text{loss}}\in R^{m\times p}$，通过求解部分数据矩阵 $\boldsymbol{Y}_{\text{loss}}$ 基于 $\boldsymbol{D}_1'\in R^{m\times k}$ 的稀疏表示，得到 $\boldsymbol{\alpha}_Y\in R^{k\times p}$。

步骤 5 利用学习后的字典 $\boldsymbol{D}_2'\in R^{m\times k}$ 与 $\boldsymbol{\alpha}_Y$ 重构出补全的完整矩阵 $\boldsymbol{Y}'_{\text{full}}\in R^{m\times p}$。

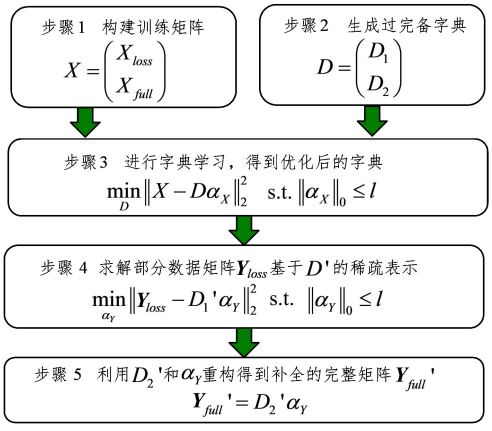


图 1 SRCM 方法框架

Fig. 1 Framework of SRCM method

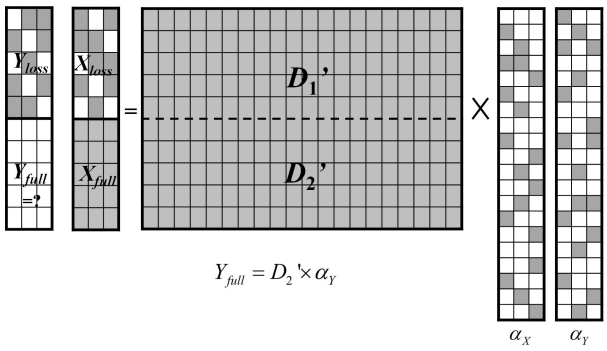


图 2 SRCM 方法的实现过程

Fig. 2 Implementation of SRCM method

下面将详细介绍 SRCM 中的几个重要步骤。

3.3 字典构造

如何确定过完备字典 $\boldsymbol{D}$，对稀疏表示来说是一个重要的研究问题，不同的过完备字典 $\boldsymbol{D}$ 可以得到不同的稀疏表示结果。由于合适的字典可以获得较低的数据补全误差，因此过完备字典 $\boldsymbol{D}$ 的创建对于 SRCM 的成功非常重要。本文的字典构造包含两个步骤：1）生成基本字典；2）进行字典学习。

3.3.1 生成基本字典

基本字典可以通过采用预先定义的数学模型得到，常见的方法包括离散傅里叶变换（Discrete Fourier Transform，DFT）、离散余弦变换（Discrete Cosine Transform，DCT）、离散小波变换（Discrete Wavelet Transform，DWT）等。此类字典构造简单，具有较高的结构性，实现较快。本文使用 DCT II 变化拼接单位对角方阵的方法生成基础字典，以保证字典的过完备性。

3.3.2 进行字典学习

基本字典的构造能够帮助 SRCM 实现电力负荷的缺失数据补全。但是，学习后的字典会获得更好的结果。字典学习的目标是通过逐步更新初始字典使其更有效地用于信号的近似。本文采用最常用的字典学习算法——K-SVD 算法。

K-SVD 算法是一种要进行 k 次迭代，且每次迭代都是用 SVD 分解的字典学习方法[22]。字典 $\boldsymbol{D}$ 的更新是逐列进行的，在每次更新中只更新一个原子和与之相应的稀疏系数，即在其他原子不变的情况下更新这个原子，直到更新完所有的原子和与之对应的稀疏系数，即完成一次迭代过程。

3.4 稀疏表示

在构造完合适的字典之后，SRCM 接下来就是对测试数据进行稀疏表示。由于 l_0 范数最小化的稀疏表示问题是 NP 难问题，因此很难在合理的时间内找到全局最优解。本文主要应用贪婪策略进行 l_0 范数最小化约束的稀疏求解。

贪婪策略是解决 NP 难问题最常见的策略。因此，大量的文献关注如何利用贪婪策略来获得稀疏表示的近似解。该策略是一种顺序选择基矢量的方法，逐步选择局部最优解，直到迭代终止。采用贪婪策略的代表性方法之一是匹配追踪（Matching Pursuit，MP）算法。但是，MP 算法存在所选择的原子不与残差正交的情况，这就导致了在该算法中，虽然每一步的迭代非常简单，但可能在收敛之前存在大量迭代的现象。本文所采用的正交匹配追踪（Orthogonal Matching Pursuit，OMP）算法克服了这一困难，使得每次选择的原子都会与之前的原子正交，且 OMP 算法通常能够得到比 MP 算法更为精确的稀疏解。

4 实验设计及结果分析

本文在澳大利亚电力负荷数据集上通过实验来说明 SRCM 的有效性。选择该数据集的原因是其从 1998 年至今一直在更新，且存在部分的数据缺失，本文的研究工作对该数据集具有一定的现实意义。

4.1 数据集简介

澳大利亚电力负荷数据集由 The Australian Energy Market Operator 提供并维护[1)]。其主要包括澳大利亚 5 个州每半小时记录一次的电力负荷数据。为了验证 SRCM 的有效性，本文选取昆士兰州、新南威尔士州与维多利亚州 3 个数据集从 2014 年至 2017 年 3 月份的电力负荷数据，该部分数据均没有缺失。本文将 2014 年至 2016 年的数据作为训练集，2017 年 1 月至 3 月的数据作为测试集。

4.2 对比算法

本文通过与线性插值、三次样条插值、Hermite 插值、基于相关性的 KNN 算法、文献[15-16]中提出的两种基于压缩感知的补全算法，以及没有经过字典学习的 SRCM 算法（Sparse Representation Completion Method-No Learning，SRCM-NL）进行性能比较。

1) http://www.aemo.com.au/Electricity/National-Electricity-Market-NEM/Data-dashboard#aggregated-data

线性插值是一种基于一维数据的插值方法，它根据一维数据序列中待插值点的邻近数据点来进行数据的估计。

三次样条插值与 Hermite 插值均是利用已知点建立插值函数，其缺失数据由对应点的函数值补全。

KNN 是一种简单的算法，但经常用于解决许多机器学习问题，包括补全缺失值。普通的 KNN 通过取矩阵中最近的 k 个邻居的平均值来补全缺失值。而基于相关性的 KNN 算法则通过使用来自其相邻行的 k 个邻居的平均值来计算出缺失值。

文献[15]提出了一种针对时空缺失数据的压缩感知补全算法。该方法利用交通数据矩阵的低秩性，根据接收到的探测数据矩阵，采用稀疏正则化矩阵分解的思想，将低阶近似与 SVD 插值结合以找到一个符合低秩要求的估计矩阵。

文献[16]则针对一维时间序列缺失数据提出了另一种基于压缩感知的缺失数据预测算法。该算法首先利用时间序列的时域平滑特性设计稀疏表示基，从而将缺失数据预测问题转化成稀疏向量恢复问题；其次，根据未缺失数据的位置特点设计了与稀疏表示基相关性低的观测矩阵，从而预测出缺失数据。

没有经过字典学习的 SRCM 算法（SRCM-NL）指仅使用 DCT II 变化拼接单位对角方阵形成的字典，不经过字典学习环节。

4.3 实验设计

电力负荷的缺失值补全结果的评估标准是通过比较缺失数据的真实值与补全值得到的。实验采用缺失补全误差率 ER 值来衡量补全的误差，如式(4)所示：

$$ER=\sum_{i,j;s(i,j)=0}\frac{|\boldsymbol{Y}'_{\text{full}}(i,j)-\boldsymbol{Y}_{\text{full}}(i,j)|}{\boldsymbol{Y}_{\text{full}}(i,j)}\times\frac{100}{c} \tag{4}$$

其中，$\boldsymbol{Y}_{\text{full}}$ 表示完整数据矩阵的实际值，$\boldsymbol{Y}'_{\text{full}}$ 表示补全的完整矩阵的补全值，c 表示缺失数据的个数，$s(i,j)=0$ 表示只计算缺失数据经过补全后与真实数据之间的误差。

对于插值法，由于要求是一维数据序列，因此采用 2014 年至 2017 年 3 月的数据作为待补全数据。其中仅设置 2017 年 1 月至 3 月的数据缺失，缺失率也是针对 2017 年 1 月至 3 月的数据而言。

对于机器学习算法而言，因澳大利亚数据集每半小时记录一次电力负荷值，即每一天都有 48 个电力负荷值，本文将每一天的数据作为完整数据矩阵 $\boldsymbol{X}_{\text{full}}$ 和 $\boldsymbol{Y}_{\text{full}}$ 的一列。其次，基于相关性的 KNN 算法的参数 k，文献[15]中时空压缩感知估计算法的参数 λ，文献[16]中时序压缩感知预测算法的参数 k 以及 SRCM-NL 和 SRCM 的参数 l 均采用在一定区间内通过网格搜索得到的最优参数。

对于采集位置矩阵，采用简单的随机缺失模式，即矩阵中的数据进行独立的随机缺失。对每个方法都设置不同的缺失率，并进行 10 次实验，最后所得的 ER 值为 10 次实验的平均值。

4.4 实验结果分析

本实验主要测试了 3 种不同的数据集在数据随机缺失模式下不同算法之间的误差率。数据缺失率控制在 5%～95%之间。表 1－表 3 与图 3－图 5 分别显示了在昆士兰州、新南威尔士州、维多利亚州的电力负荷数据集中 SRCM 与不同算法在各缺失率下的比较结果。其中，由于 3 种插值方法基本一致，因此在图 3－图 5 中，这 3 种插值方法仅选取线性插值进行比较。

表 1 昆士兰州数据集中 SRCM 与不同算法的 ER 值比较

Table 1 Comparison of ER values between SRCM and different algorithms in Queensland data set

（单位：%）

缺失率	线性插值	三次样条插值	Hermite 插值	基于相关性的 KNN	时空压缩感知估计算法	时序压缩感知预测算法	SRCM-NL	SRCM
5	20.21	20.22	20.22	10.96	1.13	8.53	5.04	0.18
15	20.72	20.73	20.73	17.13	1.14	8.49	8.32	1.00
25	20.75	20.76	20.76	17.28	1.20	8.53	13.96	1.44
35	20.64	20.65	20.65	17.48	1.26	8.80	22.93	1.53
45	20.56	20.58	20.57	17.22	1.42	8.93	33.98	1.45
55	20.61	20.62	20.62	17.10	2.06	8.93	49.20	1.50
65	20.68	20.70	20.69	17.04	5.26	9.24	61.61	1.62
75	20.55	20.56	20.56	17.24	16.15	9.42	76.34	2.03
85	20.62	20.64	20.64	17.28	26.84	9.69	100	2.03
95	20.61	20.63	20.62	17.13	25.25	10.53	100	2.14

表 2 新南威尔士州数据集中 SRCM 与不同算法的 ER 值比较

Table 2 Comparison of ER values between SRCM and different algorithms in NSW data set

（单位：%）

缺失率	线性插值	三次样条插值	Hermite 插值	基于相关性的 KNN	时空压缩感知估计算法	时序压缩感知预测算法	SRCM-NL	SRCM
5	16.20	16.22	16.22	11.08	0.67	10.08	4.15	0.11
15	16.39	16.44	16.43	10.64	0.59	9.85	7.76	0.90
25	16.45	16.48	16.48	10.29	0.62	9.83	13.72	1.99
35	16.58	16.61	16.60	11.38	0.70	9.92	16.73	2.78
45	16.59	16.62	16.62	10.35	0.88	10.14	36.12	2.16
55	16.55	16.58	16.57	10.18	1.41	10.23	49.28	2.62
65	16.45	16.49	16.48	10.80	3.03	10.22	61.45	2.17
75	16.46	16.49	16.49	10.69	9.95	10.63	74.85	3.57
85	16.48	16.51	16.51	10.84	22.18	11.09	100	4.72
95	16.51	16.54	16.54	10.84	22.52	12.07	100	3.98

表 3 维多利亚州数据集中 SRCM 与不同算法的 ER 值比较

Table 3 Comparison of ER values between SRCM and different algorithms in Victoria data set

(单位:%)

缺失率	线性插值	三次样条插值	Hermite 插值	基于相关性的 KNN	时空压缩感知估计算法	时序压缩感知预测算法	SRCM-NL	SRCM
5	23.45	23.47	23.46	9.31	0.68	9.95	3.61	0.73
15	23.08	23.11	23.10	8.22	0.70	9.84	6.38	0.96
25	23.25	23.26	23.26	7.36	0.78	9.83	11.52	1.92
35	23.30	23.31	23.31	7.58	0.86	10.00	22.17	2.23
45	23.20	23.22	23.21	7.20	1.14	10.01	37.42	3.24
55	23.30	23.32	23.31	7.09	1.78	10.01	56.91	3.20
65	23.08	23.10	23.10	7.99	3.54	10.26	61.44	2.83
75	23.22	23.24	23.23	7.66	11.05	11.37	75.91	3.04
85	23.17	23.20	23.19	7.77	23.38	11.09	100	4.30
95	23.21	23.23	23.23	8.73	23.08	12.07	100	3.78

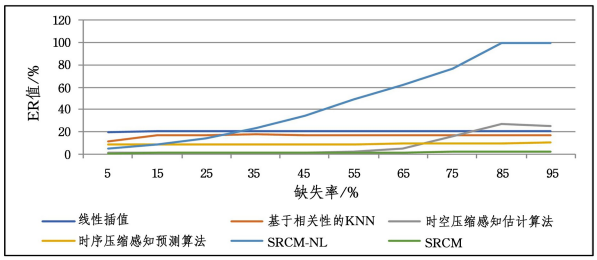


图 3 昆士兰州数据集中 SRCM 与不同算法的 ER 值比较

Fig. 3 Comparison of ER values between SRCM and different algorithms in Queensland data set

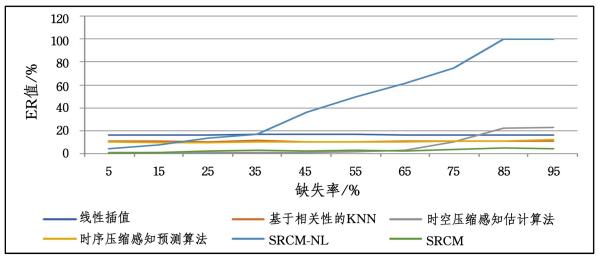


图 4 新南威尔士州数据集中 SRCM 与不同算法的 ER 值比较

Fig. 4 Comparison of ER values between SRCM and different algorithms in NSW data set

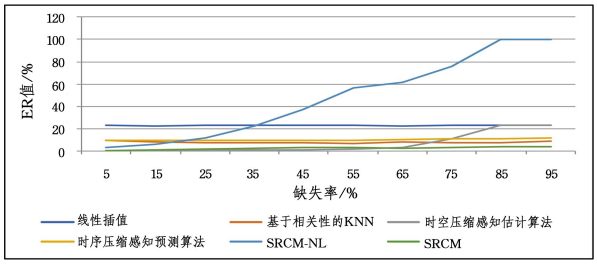


图 5 维多利亚州数据集中 SRCM 与不同算法的 ER 值比较

Fig. 5 Comparison of ER values between SRCM and different algorithms in Victoria data set

由表 1—表 3 以及图 3—图 5 可以得出:1)即使数据缺失 95%,SRCM 仍然可以重建出误差率小于 5%的电力负荷数据,在所有的对比方法中排名第一;2)3 种传统的插值方法在各种缺失率下都不如基于相关性的 KNN 算法以及时序压缩感知预测算法,与 SRCM 相比更是相差甚远;3)基于相关性的 KNN 算法以及时序压缩感知预测算法虽然在缺失率较高的情况下优于时空压缩感知估计算法以及 SRCM-NL,但是在低缺失率下与时空压缩感知估计算法以及 SRCM-NL 有很大的差距;4)时空压缩感知估计算法在缺失率较低时,其效果与 SRCM 相差不大,但是在缺失率达到 55%以上时,其误差明显增长;5)SRCM-NL 虽然在缺失率为 15%及以下时比插值法、基于相关性的 KNN 算法以及时序压缩感知预测算法的效果好,但是随着缺失率的增加,其误差高于其他所有算法,尤其当缺失率超过 85%时,SRCM-NL 已经不能准确补全缺失数据。这说明字典学习环节对于 SRCM 至关重要。

综上所述,SRCM 相比其他方法可以获得最低的补全误差,当缺失 95%的数据时,SRCM 依然可以成功实现误差率在 5%以内的数据重建。

结束语 本文主要研究了电力负荷中的数据缺失和重建问题,并提出一种基于稀疏表示的数据补全方法 SRCM。SRCM 的主要思想是利用训练集对稀疏表示的过完备字典进行学习,当过完备字典对于训练集具有较小的重构误差时,其对测试集可进行较好的缺失数据补全。通过随机模拟电力负荷数据的缺失情况,本文将 SRCM 与几种常见的补全方法进行了性能比较。实验结果表明,SRCM 算法优于其他补全方法,可以实现缺失数据的有效补全,即使对于缺失率高达 95%的数据,补全误差仍然可以在 5%以内。在未来的工作中,我们将对该方法进行进一步的研究,希望通过主动学习技术来智能地选择尽可能少的训练数据,以降低字典学习的计算成本。

参 考 文 献

[1] FAN W Q, ZHANG W, LI Y G, et al. Ultra short-term load forecasting for micro-grid based on improved human comfort index[J]. Guangdong Electric Power, 2017, 30(4): 137-142.

[2] LI H. Residual error GM(1,1) model improved by Markov method for long-term and medium-term load forecasting[J]. Shaanxi Electric Power, 2017, 45(5): 75-77.

[3] YANG H X, DENG Y J, LIU Z B, et al. Study on electric load forecasting with historical bad data[J]. Power System Protection and Control, 2017, 45(15): 62-68.

[4] CHEN Z H, ZHANG Y, WU Z G. Application of RBF neural network in medium and long-term load forecasting[J]. Proceedings of the CSU-EPSA, 2006, 18(1): 15-19.

[5] LAKSHMINARAYAN K. Imputation of missing data in industrial databases[J]. Applied Intelligence,1999,11:259-275.

[6] WU S F,CHANG C Y,LEE S J. Time series forecasting with missing values[C]// 2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom). 2015:151-156.

[7] DING Q,LI S. Research on Resampling Application of Intelligent Substation and Error Analysis of Linear Interpolation Method[J]. Power System Protection and Control, 2015, 43(23):132-136.

[8] ZHU Q W,YE L,ZHAO Y N,et al. Research on Identification and Reconstruction Method of Wind Farm Output Power Abnormal Data[J]. Power System Protection and Control,2015, 43(3):38-45.

[9] RUAN Q Z,CHEN J B,ZHU G,et al. Instantaneous test data analysis of low voltage electrical equipment based on cubic spline interpolation[J]. Low Voltage Electrical Appliance,2012(10): 27-31.

[10] TAO T Y,WANG H. Simplified wind field simulation based on Hermite interpolation[J]. Engineering Mechanics,2017,34(3): 182-188.

[11] GERHARD T,SHAHLA R. Improved methods for the imputation of missing data by nearest neighbor methods[J]. Computational Statistics and Data Analysis,2015,90:84-99.

[12] NEWSHAM G R,BIRT B J. Building-level occupancy data to improve arima-based electricity use forecasts[C]// Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building, ACM. New York, USA, 2010: 13-18.

[13] SHI W,ZHU Y,ZHANG J,et al. Improving power grid monitoring data quality: An efficient machine learning framework for missing data prediction[C]// 2015 IEEE 17th International Conference on High Performance Computing and Communications. IEEE,2015:417-422.

[14] KONG L,XIA M,LIU X Y,et al. Data loss and reconstruction in sensor networks[C]// INFOCOM. 2013:1654-1662.

[15] ZHU Y,LI Z,ZHU H,et al. A Compressive Sensing Approach to Urban Traffic Estimation with Probe Vehicles[J]. IEEE Transactions on Mobile Computing,2013,12(11):2289-2302.

[16] SONG X X,GUO Y,LI N,et al. Missing Data Prediction Based on Compressive Sensing in Time Series[J]. Computer Science, 2019,46(6):35-40.

[17] ZHANG Y,ROUGHAN M,WILINGER W,et al. Spatio-temporal compressive sensing and internet traffic matrices[J]. ACM SIGCOMM Computer Communication Review, 2009, 39(4): 267.

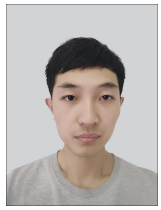
[18] MEI J L,YOHANN D C,YANNIG G,et al. Nonnegative matrix factorization with side information for time series recovery and prediction[J]. IEEE Transactions on Knowledge and Data Engineering,2018:1.

[19] WANG Z H,HORNG G J,HSU T H,et al. Heart sound signal recovery based on time series signal prediction using a recurrent neural network in the long short-term memory model[J]. The Journal of Supercomputing,2019(1):1-18.

[20] STRAUMAN A S,BIANCHI F M,MIKALSEN K Ø. Classification of postoperative surgical site infections from blood measurements with missing data using recurrent neural networks [C]// International Conference on Biomedical & Health Informatics. Las Vegas,USA,2018:307-310.

[21] ZHANG Z,XU Y,YANG J,et al. A survey of sparse representation:algorithms and applications[J]. Access IEEE, 2015, 3: 490-530.

[22] AHARON M,ELAD M,BRUCKSTEINA. K-SVD:an algorithm for designing overcomplete dictionaries for sparse representation[J]. IEEE Transactions on Signal Processing, 2006, 54(11):4311-4322.



LI Pei-guan, born in 1996, B.S. His main research interests include machine learning and so on.

HUANG Fang-wan, born in 1980, M.S., senior lecturer, is a member of China Computer Federation. Her main research interests include computational intelligence, machine learning and big data analysis.