基于压缩感知的时间序列缺失数据预测算法

宋晓祥 郭 艳 李 宁 王 萌
(陆军工程大学通信工程学院 南京 210007)

摘 要 数据缺失在时间序列采集过程中频繁发生,已经严重阻碍了精确的数据分析。然而,现有的缺失数据预测算法多是从采集到的数据中发现某种规律,从而预测缺失的数据,并不适用于缺失数据较多的情况。基于此,提出了一种基于压缩感知的缺失数据预测算法。首先,该算法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,从而保证了算法的重构性能。仿真结果表明,即使数据缺失率高达90%,所提方法依然可以非常有效地预测出缺失数据。

关键词 时间序列,缺失数据,压缩感知
中图法分类号 TN911.7 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.06.004

Missing Data Prediction Based on Compressive Sensing in Time Series

SONG Xiao-xiang GUO Yan LI Ning WANG Meng
(College of Communications Engineering, Army Engineering University, Nanjing 210007, China)

Abstract The frequent occurrence of data loss in time series acquisition process has seriously hindered the accurate data analysis. However, most of the existing methods mainly find a certain pattern from the collected data to predict the missing data, which are only feasible to be applied to the case where only a low ratio of collected data are missing. In view of the problem above, this paper proposed an algorithm of missing data prediction based on compressive sensing. The missing data prediction problem is formulated as the multiple sparse vectors recovery problem. Firstly, the sparse representation basis is designed by making use of the temporal smoothness of time series, thus transforming the missing data prediction problem into the problem of the sparse vector recovery. Secondly, the observation matrix is designed based on the location characteristics of the data that are not missing, which is lowly coherent with the designed representation bases, thus ensuring the reconstruction performance of the proposed algorithm. The simulation results show that the proposed algorithm can predict the missing data very effectively even if the ratio of data loss is as high as 90%.

Keywords Time series, Missing data, Compressive sensing

1 引言

大数据技术在互联网业务中取得初步成功后,以数据驱动的经营管理和决策制定已经在工业、商业和政府中得到广泛应用,数据质量也因此被认为是工业过程、市场经营和决策活动的关键问题[1]。然而,在数据采集过程中,由于恶劣的工作条件或一些无法控制的因素,采集到的原始时间序列中往往存在缺失数据,这使得原始数据的质量很难满足精确数据的分析需求[2]。根据文献[3]的研究,如果直接利用未经处理的采样数据,将导致41%以上的相关项目由于数据质量问题而失败。因此,如何高效地预测缺失数据,从而提高原始数据质量,已成为亟待解决的问题。

目前,已有大量的数据挖掘和统计方法致力于缺失数据的预测[4]。基于插值的方法是最简单的缺失数据预测算法,指数平滑和样条插值是数据插值的主要技术[5-6]。虽然这些方法容易实现,并且在某些特定的情况下是有效的,但是在大量数据缺失或连续数据缺失的情况下,其效果往往不尽如人意[7-8]。

基于模型化的方法,如数值分析模型、状态空间模型和随机模型,是预测缺失数据最常用的方法[9]。这类方法主要根据收集到的数据发现某种潜在的规律,从而预测缺失的数据[10]。文献[11]使用经典的自回归滑动平均模型(ARIMA)

到稿日期:2018-04-18 返修日期:2018-07-05 本文受国家自然科学基金(61571463,61371124,61472445),江苏省自然科学基金(BK20171401)资助。

宋晓祥(1993-),男,硕士生,主要研究方向为信号处理、大数据;郭 艳(1971-),女,博士,教授,博士生导师,主要研究方向为波束形成、认知无线电、无线传感器网络定位、自适应信号处理,E-mail:guoyan_1029@sina.com;李 宁(1967-),男,副教授,硕士生导师,主要研究方向为Ad hoc网络、无线认知网络;王 萌(1983-),男,硕士,主要研究方向为信号处理。

来预测缺失数据并取得了不错的效果，但是 ARIMA 模型并没有充分利用在数据缺失之后的时间内采集到的数据，从而在一定程度上影响了预测效果。文献[12]利用支持向量机(SVM)构造缺失数据预测框架，并基于此有效地预测了电网监测数据中的缺失值。但是，SVM 只适用于缺失数据较少且数据序列非常稳定的情况。文献[13]使用 BP 神经网络算法来预测缺失数据，该算法充分利用了神经网络强大的非线性拟合能力和并行处理能力。但是对于神经网络来说，通常需要一个较大的训练数据集，这在数据缺失严重的情况下是很难实现的。

基于统计学习的方法试图利用数据的统计特征确定一个特殊的概率分布，然后将最适合假定概率分布的值作为缺失的数据[14]。文献[15]使用概率主成分分析法(PPCA)来估算交通流的缺失数据，实现了良好的性能。文献[16]提出了一种基于矩阵分解的高效缺失数据预测算法。但是，常见的矩阵分解方法通常需要结合数据的内部特定特征，如空间信息、时域相关信息等，从而限制了其应用场景。文献[17]提出了一种基于时域贝叶斯网络(TBN)的动态内容矩阵分解的方法，其在预测时间序列中的缺失数据时具有良好的性能。但是作为一个概率图模型，TBN 在数据集较小时的性能较差；而当数据量较大时，TBN 的计算代价又较高。

压缩感知理论(Compressive Sensing，CS)指出，如果信号本身是稀疏的或在某个稀疏表示基下是稀疏的，就可以对信号进行欠采样，从而通过少量的采样值以大概率恢复出原信号。这与存在大量缺失数据的时间序列中仅有少量可利用的观测数据的特点十分契合。鉴于以上分析，本文提出了一种基于压缩感知的时间序列缺失数据预测算法。该算法充分利用了时间序列的时域平滑特性来设计稀疏表示基，从而将缺失数据预测问题转化为稀疏向量恢复问题。其次，该算法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，保证了算法的重构性能。为了对算法的性能进行检验，引用 3 个真实的数据集进行大量的仿真实验。结果表明，即使数据缺失率高达 90%，本文提出的算法依然可以非常有效地预测出缺失数据。

2 压缩感知理论

在压缩感知理论中，用 $x_{N\times 1}$ 表示 N 维离散信号，则：

$$x=\psi\theta \tag{1}$$

其中，$\psi_{N\times N}$ 被称为稀疏矩阵；$\theta\in R^N$ 是 x 在稀疏矩阵 ψ 下的表示系数，若 θ 中只有 $K(K\ll N)$ 个非零值，则称 θ 是 K 阶稀疏的。将稀疏的 N 维向量 x 投影到 $M(M<N)$ 维空间中，即：

$$y=\Phi x=\Phi\psi\theta=A\theta \tag{2}$$

其中，$\Phi_{M\times N}$ 被称为观测矩阵，$y\in R^M$ 表示投影后的测量向量。此时通过求解优化问题，就可以恢复出稀疏信号 $\hat{\theta}$：

$$\begin{aligned}&\theta=\arg\min\|\theta\|_0\\&\text{s.t. } y=\Phi\psi\theta\end{aligned} \tag{3}$$

最后，通过 $\hat{\theta}$ 求出原始信号 $\hat{x}$：

$$\hat{x}=\psi\hat{\theta} \tag{4}$$

3 时间序列缺失数据预测模型

为简单起见，本文只介绍一维时间序列缺失数据预测算法，多维空间可以类似扩展。设一维时间序列信号为 $x\in R^N$，若存在一个稀疏表示基 ψ 使得时间序列信号 x 稀疏，则可以通过观测矩阵 Φ 对 x 进行欠采样，并通过少量的采样值恢复出原始信号 x。

3.1 稀疏表示基的设计

众所周知，实际生活中的大多数时间序列信号都具有天然的时域平滑性，如室内温度、商品价格、城市能源消耗等，即信号 x 的值只在少数时刻发生较大变化。因此，信号 x 的两个相邻采样值之差应该只有少量较大，而其他大部分可以忽略。设时间序列 $x\in R^N$ 为 $x=\{x_1,x_2,\cdots,x_N\}$，我们考虑如式(5)所示矩阵：

$$\Omega_1=\begin{bmatrix}1&-1&0&\cdots\\0&1&-1&\cdots\\0&0&1&\cdots\\\vdots&\vdots&\vdots&\vdots\end{bmatrix} \tag{5}$$

则信号 x 在矩阵 Ω_1 下的投影向量为：

$$\theta_1=\begin{bmatrix}1&-1&0&\cdots\\0&1&-1&\cdots\\0&0&1&\cdots\\\vdots&\vdots&\vdots&\vdots\end{bmatrix}\begin{bmatrix}x_1\\x_2\\\vdots\\x_N\end{bmatrix}=\begin{bmatrix}x_1-x_2\\x_2-x_3\\\vdots\\x_N-x_1\end{bmatrix} \tag{6}$$

其中，θ_1 中的元素 x_i-x_{i+1} 表示时间序列 x 中的两个相邻采样值之差。因此，θ_1 中只有少量元素较大，而其他大部分元素可以忽略。常见的用于表示时域平滑性的矩阵还有二阶差分方程，如式(7)所示：

$$\Omega_2=\begin{bmatrix}2&-1&0&0&\cdots\\-1&2&-1&0&\cdots\\0&-1&2&-1&\cdots\\\vdots&\vdots&\vdots&\vdots&\vdots\end{bmatrix} \tag{7}$$

信号 x 在矩阵 Ω_2 下的投影向量为：

$$\theta_2=\begin{bmatrix}2&-1&0&\cdots\\-1&2&-1&\cdots\\0&-1&2&\cdots\\\vdots&\vdots&\vdots&\vdots\end{bmatrix}\begin{bmatrix}x_1\\x_2\\\vdots\\x_N\end{bmatrix}=\begin{bmatrix}2x_1-x_2\\2x_2-x_2-x_3\\\vdots\\2x_N-x_{N-1}\end{bmatrix} \tag{8}$$

令 $\psi_1=\Omega_1^{-1}$，$\psi_2=\Omega_2^{-1}$，统称 θ_1 和 θ_2 为 θ，ψ_1 和 ψ_2 为 ψ。通过利用时间序列的时域平滑性特征设计稀疏表示基 ψ，可将时间序列 x 中的缺失数据预测问题转化为恢复稀疏向量 θ。因为一旦求出 θ，就可以利用式(4)恢复出原始时间序列。

3.2 观测矩阵的设计

存在缺失数据的时间序列中仅有部分采样数据，缺失数据预测问题就是利用这些未缺失的数据恢复出原始的时间序列。为此，我们根据未缺失数据的位置设计观测矩阵 Φ，并将它们作为观测值来恢复 θ。

$$\Phi=\begin{bmatrix}1&0&0&0&\cdots\\0&0&1&0&\cdots\\0&0&0&1&\cdots\\\vdots&\vdots&\vdots&\vdots&\vdots\end{bmatrix} \tag{9}$$

设共有 M 个观测值，则 $\Phi\in R^{M\times N}$。如果矩阵 Φ 中(m,n)处的值为 1，则表明第 m 个测量值是在第 n 个采样时刻得到的。通过设计观测矩阵 Φ 和稀疏表示基 ψ，得到 $y=\Phi\psi\theta=A\theta$，$y\in R^M$。利用已知的 y 和 A，通过优化算法就可以求解 θ，从而恢复出原信号。

4　性能分析

利用优化算法求解 θ，从而实现信号重构，需要满足以下两个条件：1）信号在稀疏表示基下的表示向量足够稀疏；2）稀疏表示基与观测矩阵不相关。下面从这两个角度来分析所设计的稀疏表示基和观测矩阵的性能。

4.1　稀疏表示基的稀疏信号能力

现实生活中，信号 x 在稀疏表示基 ψ 下的稀疏表示向量并非完全是稀疏的，而是可压缩的，即稀疏表示向量中只包含一些非常大的元素，而其他的大部分元素可以被忽略。基于此，以$(\sum_{i=1}^{K}\theta_i^2)/(\sum_{i=1}^{N}\theta_i^2)$为衡量指标来分析稀疏表示基的稀疏信号能力。其中，$\theta_i$ 表示稀疏向量 θ 中的第 i 大元素；$\sum_{i=1}^{K}\theta_i^2$ 表示稀疏向量 θ 中 K 个最大元素的能量；$\sum_{i=1}^{N}\theta_i^2$ 表示稀疏向量 θ 的总能量。对于给定的 K，$(\sum_{i=1}^{K}\theta_i^2)/(\sum_{i=1}^{N}\theta_i^2)$越大，稀疏表示基的稀疏信号能力就越强。本文利用如下 3 个真实数据集中的数据来检验设计的稀疏表示基 ψ_1 和 ψ_2 的性能。

1）GSA 数据集中的数据是在加州大学圣地亚哥分校生物电路研究所化学信号实验室的一个气体传递平台上收集的。GSA 包含了从 16 个在不同浓度的空气中暴露于乙烯的化学传感器中测得的时间序列数据[18]。

2）NCSU 数据集包含了来自 NCSU 大学 32 位大学生的运动轨迹数据，这些数据由 40 个传感器每 10 s 采样一次收集而来[19]。

3）NY 数据集是一个小数据集，包含了来自纽约市 12 位志愿者的运动轨迹数据，这些数据由 40 个传感器每 10 s 采样一次收集而来[19]。

为方便比较，本文设置每个数据集的数据长度为 1 000。如图 1－图 3 所示，总体而言，稀疏表示基 ψ_1 稀疏信号的能力略强于 ψ_2。可以发现，当取 ψ_1 作为稀疏表示基时，前 50 个大元素占总能量的比重为 92%～99%；当取 ψ_2 作为稀疏表示基时，前 50 个大元素占总能量的比重为 71%～99%。无论选择哪一个稀疏表示基，稀疏向量 θ 绝大部分的能量都集中在前 50 个元素之中，即信号 x 在稀疏表示基下的稀疏表示向量足够稀疏。因此，我们有理由相信，当观测数据的数目大于 50 时，观测数量 M 就大于稀疏向量的稀疏度 K。

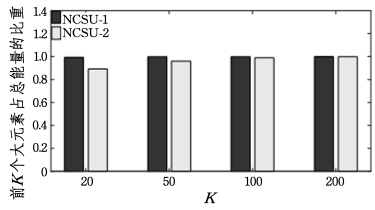

图 1　NCSU 中稀疏表示基稀疏信号的能力

Fig. 1　Ability of sparse representation basis sparse signal in NCSU

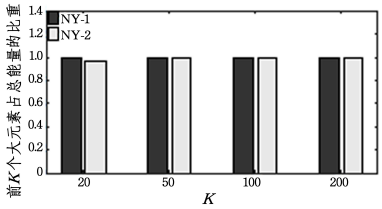

图 2　NY 中稀疏表示基稀疏信号的能力

Fig. 2　Ability of sparse representation basis sparse signal in NY

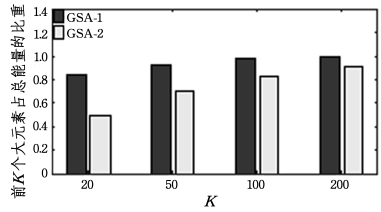

图 3　GSA 中稀疏表示基稀疏信号的能力

Fig. 3　Ability of sparse representation basis sparse signal in GSA

4.2　稀疏表示基和观测矩阵的低相关性

对于 N 维空间的一组正交基(Φ,ψ)，它们的相关性可以表示为：

$$\mu(\Phi,\psi)=\sqrt{N}\max|\langle\varphi_i^r,\psi_j\rangle|\in[1,\sqrt{N}] \tag{10}$$

其中，φ_i^r 和 ψ_j 分别表示 Φ 和 ψ 的行向量和列向量。但是，本文设计的稀疏表示基和观测矩阵都是非常稀疏的，不满足正交的条件，因此不能直接使用式(10)来计算它们之间的相关性。基于此，采用文献[20]提出的非相关性计算理论来计算它们之间的非相关性，从而间接反映相关性的大小。

对于给定的(φ,ψ)，它们之间的非相关性被定义为：

$$I(\Phi,\psi)=\min_{1\leqslant i\leqslant M}\|\theta_i\|_0 \tag{11}$$

其中，θ_i 表示矩阵 Φ 的第 i 个行向量在由稀疏表示基 ψ 各列张成的空间中的投影向量，即：

$$\theta_i=(\psi^{\mathrm{T}}\psi)^{-1}\psi^{\mathrm{T}}\varphi_i^r \tag{12}$$

其中，φ_i^r 表示观测矩阵 Φ 的第 i 个行向量。$I(\Phi,\psi)$越大，稀疏表示基 ψ 和观测矩阵 Φ 之间的非相关性就越大，相关性也就越小。

表 1 列出了当 N 等于 1000，800，500，200 时不同的稀疏表示基和观测矩阵组合之间的非相关性。可以看出，当 N 较小时，稀疏表示基 ψ_1 和 ψ_2 与观测矩阵 Φ 的非相关性大小相差不大。但是本文所使用的每个数据集的长度均为 1 000，此时稀疏表示基 ψ_2 和观测矩阵 Φ 的非相关性明显更大，即相关性更小。考虑到图 1 中$K\geqslant 50$ 时，两者的稀疏信号能力相当，本文选择 ψ_2 作为稀疏表示基。

表 1 稀疏表示基与观测矩阵的非相关性

Table 1 Non-correlation between sparse representation basis and measurement matrix

N	$N(\psi_1,\Phi)$	$N(\psi_2,\Phi)$
1000	926	998
800	743	790
500	500	498
200	200	199

5 仿真结果与分析

本节使用上文所介绍的数据集进行仿真。根据不同的数据缺失率,从完整的数据集中随机删除一些数据来模拟缺失的数据。缺失率定义为缺失数据的数量与数据总量的比值。本文采用均方根误差(RMSE)和平均运行时间(ART)作为性能评价标准。RMSE 是一种常用的衡量标准,本文中 RMSE 表示预测值与真实观测值之间的样本偏差,定义如下:

$$RMSE=\frac{1}{N}\sqrt{\sum_{ij}(x_{ij}-\hat{x}_{ij})^2} \tag{13}$$

其中,N 表示数据集的总长度,x_i 表示实际值,$\hat{x}_i$ 表示相应的预测值。

为了分析不同方法的计算复杂度,重复执行每种方法 50 次,并计算其平均运行时间(ART)。

$$ART=\frac{T}{50} \tag{14}$$

5.1 恢复算法的比较

通过设计相应的稀疏表示基和观测矩阵,本文把时间序列缺失数据预测问题转换为稀疏向量恢复问题。已有的解决稀疏向量恢复问题的算法较多,本文选取的对比算法包括基追踪(Basis Pursuit,BP)[21]、正交匹配追踪(Orthogonal Matching Pursuit,OMP)[22]和基于稀疏贝叶斯学习的改进算法(TMSBL)[23]。

图 4—图 6 给出了在各数据集中使用不同的恢复算法产生的均方根误差。

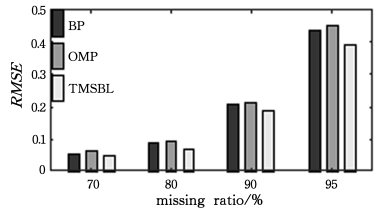

图 4 各恢复算法在 NCSU 中的性能比较

Fig. 4 Performance comparison of different recovery algorithms in NCSU

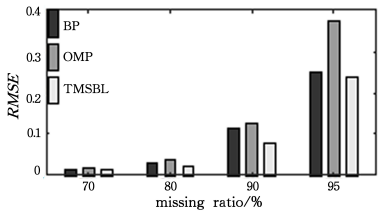

图 5 各恢复算法在 NY 中的性能比较

Fig. 5 Performance comparison of different recovery algorithms in NY

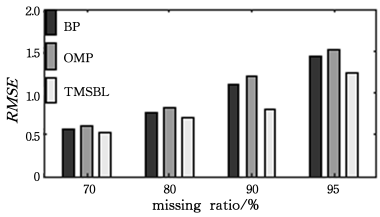

图 6 各恢复算法在 GSA 中的性能比较

Fig. 6 Performance comparison of different recovery algorithms in GSA

从图中可以看出,就均方根误差而言,无论在哪个数据集中,TMSBL 算法的性能总是最优的,BP 次之。表 2 列出了 NCSU 数据集中 3 种恢复算法在不同数据缺失率下的平均运行时间。从表中可以看出,OMP 算法所需的运行时间最短;而 TMSBL 算法的误差虽然最小,但是由于计算过程中需要将数据扩展到多维空间进行反复迭代,并且有大量的矩阵求逆运算,因此计算时间最长[24]。为了客观地分析所提方法的性能,本文使用 BP 作为恢复算法,构成 CS-BP 算法。

表 2 各恢复算法平均运行时间的比较

Table 2 Comparison of average running time of different recovery algorithms

(单位:s)

算法	数据缺失率			
	70%	80%	90%	95%
BP	0.7021	0.4985	0.3445	0.3025
OMP	0.4867	0.2757	0.1895	0.1578
TMSBL	1.9402	1.1025	0.7156	0.6117

5.2 CS-BP 算法的性能分析

为了对基于压缩感知的缺失数据预测算法(CS-BP)的性能进行有效评估,本文将其与其他 3 种算法进行仿真比较。对比算法包括:基于插值的样条插值算法(SI)、基于模型的神经网络算法(NN)、基于统计学习的概率矩阵分解算法(PMF)。

图 7—图 9 分别给出了各方法在 NCSU,NY 和 GSA 数据集上的仿真结果。

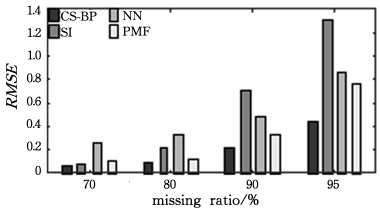

图 7 NCSU 中各算法性能的比较

Fig. 7 Performance comparison of proposed methods and other methods in NCSU

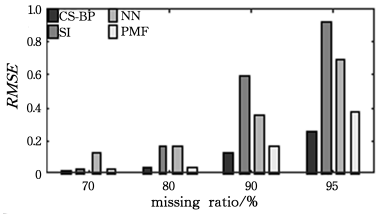

图 8 NY 中各算法性能的比较

Fig. 8 Performance comparison of proposed methods and other methods in NY

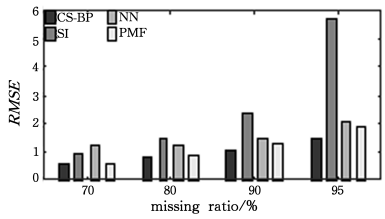

图 9　GSA 中各算法性能的比较

Fig. 9　Performance comparison of proposed methods and other methods in GSA

从图中可以看出，不管在哪个数据集上，CS-BP 算法都展示了最好的性能，说明使用该方法解决时间序列缺失数据预测问题时是适用并且非常有效的。这是因为本文充分利用了时间序列的时域平滑特性来设计稀疏表示基，从而将时间序列缺失数据的预测问题转换成了稀疏向量恢复问题。该算法与完备的压缩感知理论相契合，使其仅仅利用很少的观测值就可以精确地恢复出原始时间序列。PMF 算法从概率统计的角度解决缺失数据的预测问题，在目标函数设计中同样考虑了时间序列的时域平滑特性，因此预测误差较小。SI 算法简单地对数据进行插值运算，因此当缺失数据较多时误差最大。NN 算法是从观测到的数据中学习一种模型，然后应用该模型预测缺失的数据，当数据缺失较多时精确模型的确定变得非常困难。

在缺失率为 80% 的条件下，以平均运行时间为标准比较各种算法的计算复杂度，结果如表 3 所列。可以发现，SI 算法的平均运行时间最短，这是因为它仅仅是对数据进行简单的插值运算，最适用于实时性要求高的场合。而本文提出的 CS-BP 算法是针对一维时间序列设计的缺失数据预测模型，当数据集维数较小时，该算法的平均运行时间较短；但是随着数据维数的增多，平均运行时间因数据维数的线性叠加而变得相对较长。虽然预测精度较高，但本文算法并不适用于数据集大、实时性要求高的场合。

表 3　各算法平均运行时间的比较

Table 3　Comparison of average running time of proposed method and other methods

（单位：s）

数据集	CS-BP	PMF	NN	SI
NCSU	0.4985	1.0362	0.4852	0.1771
NY	0.5453	0.8607	0.5628	0.1719
GSA	4.7429	6.0830	2.7339	1.3626

由图 1－图 3 可知，不同数据集中的数据在同一稀疏表示基下的稀疏向量的稀疏程度是不同的。图 10 比较了稀疏向量的稀疏性对算法性能的影响。从图中可以看出，本文提出的算法在 NY 数据集上的表现优于在其他数据集上的表现。结合图 1－图 3 可以发现，NY 数据集中的数据在稀疏表示基下的稀疏向量的稀疏性最好，即当 K 相同时，$(\sum_{i=1}^{K}\theta_i^2)/(\sum_{i=1}^{N}\theta_i^2)$ 最大。因此我们有理由相信，本文算法的性能与稀疏向量的稀疏性有关，稀疏向量越稀疏，算法的性能便越好。

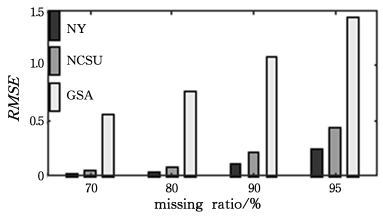

图 10　稀疏向量的稀疏性对算法性能的影响

Fig. 10　Effect of sparse vector on algorithm performance

图 11 给出了在不同数据集中分别使用 ψ_1 和 ψ_2 作为稀疏表示基时产生的均方根误差。从图中可以看出，无论在哪个数据集中，当选用 ψ_2 作为稀疏表示基时算法都具有更好的性能，这也证明了本文选择 ψ_2 作为稀疏表示基的正确性。回顾图 1－图 3 和表 1 可以发现，稀疏表示基稀疏信号的能力和稀疏表示基与观测矩阵之间的相关性共同影响着 CS-BP 算法的性能，因此在设计算法过程中要综合考虑这两个因素。

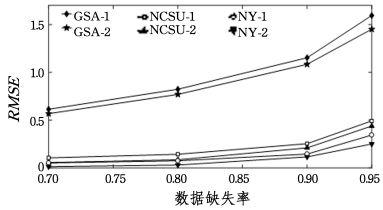

图 11　稀疏表示基的选择对算法性能的影响

Fig. 11　Effect of sparse representation basis on algorithm performance

结束语　本文提出了一种基于压缩感知的时间序列缺失数据预测算法。该算法充分利用了时间序列的时域平滑性来设计稀疏表示基，从而将缺失数据预测问题转化为稀疏向量恢复问题。此外，本文根据未缺失数据的位置特点设计了容易实现且与稀疏表示基相关性低的观测矩阵，建立了压缩感知模型。通过对比分析，本文选择 BP 作为稀疏恢复算法，提出了 CS-BP 缺失数据预测算法。仿真结果表明，即使数据缺失率高达 90%，本文所提算法依然可以非常有效地预测出缺失数据。尽管所提算法在预测精度上表现出了优越的性能，但是作为一维时间序列缺失数据预测算法，其在处理高维数据时只能将其拆分成单维时间序列来处理，平均运行时间较长。这主要是因为目前存在的稀疏恢复算法并不适用于同时恢复多个稀疏向量的情况。我们下一步的研究方向是设计一种普适性的多稀疏向量同时恢复算法，从而同时预测多个时间序列中的缺失数据。

参 考 文 献

[1] SHI W, ZHU Y, ZHANG J, et al. Improving Power Grid Monitoring Data Quality: An Efficient Machine Learning Framework for Missing Data Prediction [C] // IEEE International Conference on High Performance Computing and Communications. IEEE, 2015: 417-422.

[2] BATINI C, CAPPIELLO C, FRANCALANCI C, et al. Methodologies for data quality assessment and improvement [J]. Acm Computing Surveys, 2009, 41(3): 1-52.

[3] LUEBBERS D,GRIMMER U,JARKE M. Systematic Development of Data Mining-Based Data Quality Tools[C]// Proceedings of the 29th VLDB Conference. Morgan Kaufmann: San Francisco,2003:548-559.

[4] 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.

[5] BALOUJI E,SALOR Q,ERMIS M. Exponential smoothing of multiple reference frame components with GPUs for real-time detection of time-varying harmonics and interharmonics of EAF currents [C] // IEEE Industry Applications Society Meeting. IEEE,2017:1-8.

[6] KOZERA R,WILKOLAZKA M. Natural spline interpolation and exponential parameterization for length estimation of curves [C] // International Conference of Numerical Analysis & Applied Mathematics. AIP Publishing LLC,2017:1-140.

[7] JUNNINEN H,NISKA H,TUPPURAINEN K,et al. Methods for imputation of missing values in air quality data sets[J]. Atmospheric Environment,2004,38(18):2895-2907.

[8] HONG S T,CHANG J W. A New Data Filtering Scheme Based on Statistical Data Analysis for Monitoring Systems in Wireless Sensor Networks[C]// IEEE International Conference on High Performance Computing and Communications. IEEE,2011:635-640.

[9] FUNG D S. Methods for the estimation of missing values in time series[J/OL]. Theses Doctoratos & Masters,2006. http://ro.ecu.edu.au/theses/63.

[10] LAO W,WANG Y,PENG C,et al. Time series forecasting via weighted combination of trend and seasonality respectively with linearly declining increments and multiple sine functions[C]// 2014 International Joint Conference on Neural Networks (IJCNN). 2014:832-837.

[11] 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.

[12] 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.

[13] WEI G,KUN N,MAN C,et al. A data prediction algorithm based on BP neural network in telecom industry[C]// 2011 International Conference on Computer Science and Service System (CSSS). 2011.

[14] LI L,LI Y,LI Z. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence [J]. Transportation Research Part C,2013,34(9):108-120.

[15] QU L,LI L,ZHANG Y,et al. PPCA-based missing data imputation for traffic flow volume: a systematical approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2009, 10(3):512-522.

[16] SHI W,ZHU Y,YU P,et al. Effective Prediction of Missing Data on Apache Spark over Multivariable Time Series[J]. IEEE Transactions on Big Data,2017,PP(99):1.

[17] CAI Y,TONG H,FAN W,et al. Fast mining of a network of coevolving time series[C]// The 2015 SIAM International Conference on Data Mining. 2015:298-306.

[18] FONOLLOSA J,SHEIK S,HUERTA R,et al. Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring[J]. Sensors & Actuators,2015,215:618-629.

[19] RHEE I,SHIN M. Mobility traces[OL]. http://carwdad.org/ncsu/mobilitymodels.

[20] WU X,LIU M. In-situ soil moisture sensing:Measurement scheduling and estimation using compressive sensing [C]// Proceedings of the 11th ACM International Conference on Information Processing in Sensor Networks. IEEE,2012:1-12.

[21] CHEN S S,DONOHO D L,SAUNDERS M A. Atomic decomposition by basis pursuit[J]. SIAM Review,2001,43(1):129-159.

[22] TROPP J A,GILBERT A C. Signal recovery from random measurements via orthogonal matching pursuit[J]. IEEE Transactions Information Theory,2007,53(12):4655-4666.

[23] ZHANG Z,RAO B D. Sparse Signal Recovery With Temporally Correlated Source Vectors Using Sparse Bayesian Learning [J]. IEEE Journal of Selected Topics in Signal Processing, 2011, 5(5):912-926.

[24] Al-SHOUKAIRI M,SCHNITER P,RAO B D. A GAMP Based Low Complexity Sparse Bayesian Learning Algorithm [J]. IEEE Transactions on Signal Processing,2018,66(2):294-308.