基于稀疏贝叶斯学习的协同进化时间序列缺失数据预测算法

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

摘 要 针对大多数已有算法在预测协同进化时间序列中的缺失数据时只适用于缺失数据较少情况的问题,提出了一种高效的缺失数据预测算法。首先,应用压缩感知理论,将协同进化时间序列中的缺失数据预测问题建模成多稀疏向量恢复问题;其次,从稀疏表示向量是否足够稀疏和感知矩阵是否满足有限等距特性两个方面分析了模型的性能;最后,针对协同进化时间序列的特点设计了一种基于稀疏贝叶斯学习的高效恢复算法,该算法可以通过学习得到部分支持信息,从而同时解决多个稀疏向量的恢复问题。仿真结果表明,所提算法可以同时有效地预测出多个时间序列中的缺失数据。

关键词 协同进化时间序列,缺失数据,稀疏表示向量,感知矩阵,稀疏贝叶斯学习
中图法分类号 TN911.7 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.07.033

Missing Data Prediction Algorithm Based on Sparse Bayesian Learning in Coevolving Time Series

SONG Xiao-xiang GUO Yan LI Ning YU Dong-ping
(College of Communications Engineering, Army Engineering University of PLA, Nanjing 210007, China)

Abstract In view of most of the existing algorithms in predicting the missing data in the coevolving time series are only feasible to be applied to the case where only a low ratio of collected data are missing, an efficient missing data prediction method was proposed in this paper. Firstly, the compressive sensing theory is applied to model the missing data prediction problem in the coevolving time series to the problem of multiple sparse vectors recovery. Secondly, the validity of the model is analyzed from two aspects: whether the sparse representation vector is sufficiently sparse and the sensing matrix satisfies the restricted isometry property. Finally, the novel recovery algorithm based on sparse Bayesian learning, which can solve multiple sparse vector recovery problems by learning some support information, is designed for the characteristics of coevolving time series. Simulation results show that the proposed algorithm can effectively predict the missing data in multiple time series simultaneously.

Keywords Coevolving time series, Missing data, Sparse representation vector, Sensing matrix, Sparse bayesian learning

1 引言

随着互联网的普及和智能终端数量的增加,近几年全球产生的数据约占迄今为止全球数据总量的90%[1]。在现实生活中,有很多应用都包含多个时间序列[2]。例如,在个人卫生保健系统中,个人的健康状况需要通过检测多种时间信号(如心率、血压和睡眠质量)来知晓,众所周知,血压通常会随着心跳的增加而升高,因此当血压值发生较大变化时,心跳监测器的值也会发生相应变化;又例如,在一个装有各种传感器的智能空间中,湿度传感器的测量值总是随着温度的上升而变大。本文中就称这样的多时间序列为协同进化时间序列。

协同进化时间序列的普遍存在为以数据为驱动的产业创造了巨大的机遇。然而,在数据采集过程中,恶劣的工作条件或一些无法控制的因素,往往使得采集到的原始时间序列中存在缺失数据。目前,大部分数据分析处理方法都要求数据集是完整的,而数据缺失则会影响数据本应该反映的真实信息以及对其的进一步开发利用[3]。

近年来,很多学者已经致力于解决数据缺失问题。基于插值的方法通过基于观测数据的插值来预测缺失的数据,指数平滑和样条是数据插值的主要技术[4-5]。AL-DEEK等[6]建立了各种回归模型来估计缺失的数据,并通过比较得出二次回归模型具有最好的预测效果。Boyles等[7]使用3种类型的缺失数据,对各种回归模型的预测精度进行了评价。对比结果表明,这些回归模型虽然易于开发和应用,但只适用于特定的条件。文献[8]利用季节内核来度量时间序列之间的相似性,并联合卡尔曼滤波提出了季节性自回归滑动平均模型

到稿日期:2018-05-27 返修日期:2018-07-04 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(通信作者);**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;**余东平**(1989-),男,博士,主要研究方向为无线传感器网络定位、信号处理。

(Seasonal Autoregressive Integrated Moving Average Model, SARIMA);但 SARIMA 模型旨在预测具有潜在季节性的数据,因此如果时间序列没有严格的内部季节性,将很难对数据建模。文献[9]使用内核概率主成分分析法(Kernel Probabilistic Principle Component Analysis,KPPCA)来预测交通矩阵中的缺失数据,并证明了相比于概率主成分分析法(Probabilistic Principle Component Analysis,PPCA),KPPCA 表现出了更好的预测性能。文献[10]利用基于时域动态矩阵分解(Temporal Dynamic Matrix Factorization,TDMF)的方法来预测多变量时间序列中的缺失数据;然而,TDMF 的性能很容易受到参数的影响,且对理想参数的设定至今没有科学的理论作为指导。文献[11]探讨了基于递归神经网络(Recurrent Neural Network,RNN)来处理缺失数据的策略;然而,RNN 需要一个较大的训练数据集,因此当有大量数据缺失时,其很难发现数据的潜在规律。Cai 等[12]基于时域贝叶斯网络(Temporal Bayesian Networks,TBN)提出了一种动态内容矩阵分解方法,该方法在预测多源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 在数据集较小的情况下性能较差,而当数据量较大时其计算代价又很大。

针对上述问题,本文提出了一种高效的缺失数据预测算法。该算法充分利用时间序列的时域平滑特性设计了稀疏表示基,并根据未缺失数据的位置特点设计了相应的观测矩阵,从而基于压缩感知理论将协同进化时间序列中的缺失数据预测问题建模成了多稀疏向量恢复问题。但若要精确地求解稀疏向量恢复问题,则必须满足两个条件:1)稀疏表示向量必须足够稀疏;2)感知矩阵必须满足有限等距特性。因此,本文从这两个方面出发对模型的性能进行了分析。针对已有的算法在解决多稀疏向量同时恢复问题时需要事先知晓多个稀疏向量的共同支持信息的情况,充分利用了协同进化时间序列的特点,设计了一种基于稀疏贝叶斯学习的多稀疏向量恢复算法。该算法可以通过学习得到部分支持信息,从而同时解决多个稀疏向量的恢复问题。为了验证算法的性能,本文引用了 3 个真实的数据集进行了大量的仿真。结果表明,所提算法在预测误差和平均运行时间上具有优越的性能。

2 系统模型

为了更清楚地说明问题,表 1 列出了一个存在缺失数据的协同进化时间序列。

表 1 存在缺失数据的协同进化时间序列

Table 1 Coevolving time series with missing data

	s_1	s_2	s_3	s_4	…	s_K
t_1	0.2	?	?	0.3	…	1.9
t_2	?	0.4	0.5	?	…	?
t_3	?	0.7	?	0.6	…	?
t_4	0.8	?	?	?	…	2.1
…	…	…	…	…	…	…
t_N	?	0.1	1.2	1.3	…	?

表 1 中,$S=\{s_1,s_2,s_3,s_4,\cdots,s_K\}$ 表示多源时间序列,$s_j\in R^N$ 表示从第 j 个数据源收集到的数据,t_i 表示第 i 个采样时刻,s_{ij} 表示第 i 个采样时刻在第 j 个数据源处的采样值,"?"表示缺失数据。此外,我们还定义了一个矩阵 $W\in R^{N\times K}$ 来表示 S 中的数据是否缺失。

$$W_{ij}=\begin{cases}0, & \text{如果 } s_{ij} \text{ 缺失}\\ 1, & \text{否则}\end{cases} \tag{1}$$

本文工作旨在利用全部或者部分观测数据来预测所有缺失的数据。

2.1 稀疏表示基的设计

压缩感知理论表明,如果信号 $s\in R^N$ 稀疏,即 $\|s\|_0\ll N$,则能够按照观测矩阵 $\psi\in R^{M\times N}$,以低于奈奎斯特定律的速率对其采样,并通过观测值 $y_{M\times 1}=\psi s$ 以高概率恢复出原信号。实际中的很多信号虽然本身并不是稀疏的,但是能在某个稀疏表示基 φ 下稀疏表示,即 $s=\varphi x$,$\|x\|_0\ll N$,同样可以按照观测矩阵 ψ 对其进行欠采样,并通过观测值 $y=\psi\varphi x$ 以高概率恢复出原始信号[13]。实际上,大多数时间序列信号都具有天然的时域平滑性,比如室内的温度、城市的能源消耗、商品的价格等都只在少数时刻才会发生较大变化。因此,信号 s 的两个相邻采样值的差值中应该只有少量值较大,而其他大部分可以被忽略。因此,设计如式(2)所示的矩阵:

$$\Omega_1=\begin{bmatrix}1 & -1 & 0 & \cdots\\ 0 & 1 & -1 & \cdots\\ 0 & 0 & 1 & \cdots\\ \vdots & \vdots & \vdots & \vdots\end{bmatrix} \tag{2}$$

设 j 个数据源收集到的数据可以表示为:$s_j=\{s_{1j},s_{2j},s_{3j},\cdots,s_{Nj}\}$,$s_j$ 在矩阵 Ω_1 下的投影向量为:

$$x_{j1}=\begin{bmatrix}1 & -1 & 0 & \cdots\\ 0 & 1 & -1 & \cdots\\ 0 & 0 & 1 & \cdots\\ \vdots & \vdots & \vdots & \vdots\end{bmatrix}\begin{bmatrix}s_{1j}\\ s_{2j}\\ \vdots\\ s_{Nj}\end{bmatrix}=\begin{bmatrix}s_{1j}-s_{2j}\\ s_{2j}-s_{3j}\\ \vdots\\ s_{Nj}-s_{1j}\end{bmatrix} \tag{3}$$

x_{j1} 中的元素 $s_{ij}-s_{(i+1)j}$ 表示时间序列 s_j 的两个相邻采样值之差。因此,x_{j1} 中只有少量元素较大,其他大部分元素可以被忽略。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(4)所示:

$$\Omega_2=\begin{bmatrix}2 & -1 & 0 & 0 & \cdots\\ -1 & 2 & -1 & 0 & \cdots\\ 0 & -1 & 2 & -1 & \cdots\\ \vdots & \vdots & \vdots & \vdots & \vdots\end{bmatrix} \tag{4}$$

s_j 在矩阵 Ω_2 下的投影向量为:

$$x_{j2}=\begin{bmatrix}2 & -1 & 0 & 0 & \cdots\\ -1 & 2 & -1 & 0 & \cdots\\ 0 & -1 & 2 & -1 & \cdots\\ \vdots & \vdots & \vdots & \vdots & \vdots\end{bmatrix}\begin{bmatrix}s_{1j}\\ s_{2j}\\ \vdots\\ s_{Nj}\end{bmatrix}=\begin{bmatrix}2s_{1j}-s_{2j}\\ 2s_{2j}-s_{1j}-s_{3j}\\ \vdots\\ 2s_{Nj}-s_{(N-1)j}\end{bmatrix} \tag{5}$$

x_{j2} 中的元素 $(s_{ij}-s_{(i-1)j})-(s_{(i+1)j}-s_{ij})$ 近似表示时间序列 s_j 变化的加速度的大小。因此,x_{j2} 中也只有少量元素较大,其他大部分元素可以被忽略。令 $\psi_1=\Omega_1^{-1}$,$\psi_2=\Omega_2^{-1}$,统称

Ω_1 和 Ω_2 为 Ω，x_{j1} 和 x_{j2} 为 x_j，ψ_1 和 ψ_2 为 ψ，则协同进化时间序列可以表示为 $s_j=\psi x_j(j=1,2,3,\cdots,K)$。

2.2　观测矩阵的设计

如上所述，矩阵 W 的第 j 列表示时间序列 s_j 中的数据是否缺失。本文将利用那些未缺失的数据作为测量值来恢复原时间序列。因此，观测矩阵的设计应与未缺失数据的位置逐一对应。基于此，设计如式(6)所示的观测矩阵：

$$\psi_j=\begin{bmatrix}1&0&0&0&\cdots\\0&0&1&0&\cdots\\0&0&0&1&\cdots\\\vdots&\vdots&\vdots&\vdots&\vdots\end{bmatrix} \tag{6}$$

如果矩阵 $\psi_j\in R^{m_j\times N}$ 中 (m,n) 处的值为 1，则表明第 m 个测量值是在第 n 个采样时刻获得的。令 $y_j\in R^{m_j}$ 表示 s_j 中未缺失数据的集合，则有：

$$y_j=\psi_j\varphi x_j=A_j x_j \tag{7}$$

现要解决的问题即是通过 y_j 和 A_j 求 x_j，而理想情况下 x_j 是稀疏的。因此，本文通过设计相应的稀疏表示基和观测矩阵，将共进化时间序列中的缺失数据预测问题建模成了多稀疏向量恢复问题。

3　性能分析

根据压缩感知理论[14]，要实现信号重构，则必须满足以下两个条件：1)时间序列在稀疏表示基下的稀疏向量足够稀疏；2)感知矩阵满足有限等距特性。下文将从这两个角度来验证所设计模型的性能。

3.1　稀疏表示基的稀疏信号能力

压缩感知理论指出，只要稀疏表示向量只包含一些非常大的元素，而其他元素可以忽略，即可认为其是可压缩的。基于此，以 $\sum_{i=1}^{n}(x_j^i)^2/\sum_{i=1}^{N}(x_j^i)^2$ 为标准来衡量稀疏表示基的稀疏信号能力。其中，x_j^i 表示稀疏向量 x_j 中第 i 大的元素；$\sum_{i=1}^{n}(x_j^i)^2$ 表示 x_j 中 n 个最大元素的能量；$\sum_{i=1}^{N}(x_j^i)^2$ 表示向量 x_j 的总能量。对于给定的 n，$\sum_{i=1}^{n}(x_j^i)^2/\sum_{i=1}^{N}(x_j^i)^2$ 越大，稀疏表示基的稀疏信号能力就越强。接下来，使用下文数据集中的数据来检验所设计的一阶稀疏矩阵(First Derivative Approximation，FD)和二阶稀疏矩阵(Second Derivative Approximation，SD)的性能。

1)KAIST 数据集是一个小数据集，包含了来自 KAIST 大学生的运动轨迹数据。这些数据是由 40 个传感器每 10 s 采样一次收集而来[15]。

2)MOTES 数据集是一个中等数据集，包含了在英特尔伯克利研究实验室部署的 54 个传感器历时一个月采集而来的时间序列数据，这些数据是由 54 个传感器每 31 s 采样一次收集而来[16]。

3)GSA 数据集是一个大数据集，数据集的气体传感器阵列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[17]。实验中，16 个化学传感器被置于含有不同浓度的乙烯的空间中，GSA 数据集就是由这 16 个传感器测得的时间序列数据组成的。

图 1－图 3 给出了稀疏表示基在这 3 个数据集中的稀疏信号能力。

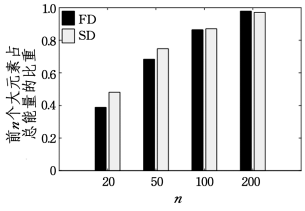

图 1　KAIST 中稀疏表示基的稀疏信号能力

Fig. 1　Ability of sparse representation basis in KAIST

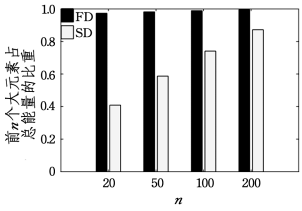

图 2　MOTES 中稀疏表示基的稀疏信号能力

Fig. 2　Ability of sparse representation basis in MOTES

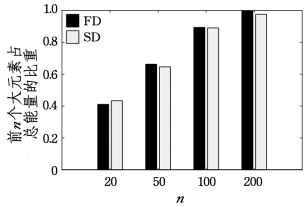

图 3　GSA 中稀疏表示基的稀疏信号能力

Fig. 3　Sparse signal ability of sparse representation basis in GSA

在每个数据集中，我们都比较了当 n 为 20，50，100，200 时，稀疏表示矩阵分别为一阶差分矩阵 Ω_1 和二阶差分矩阵 Ω_2 时 $\sum_{i=1}^{n}(x_j^i)^2/\sum_{i=1}^{N}(x_j^i)^2$ 的大小。从图 1－图 3 中可以看出，在不同的数据集中，Ω_1 和 Ω_2 的稀疏能力是不同的。通过观察比较可以发现，当采用 Ω_1 作为稀疏表示矩阵时，最大的 50 个元素占据了信号总能量的 64.6%～98.2%，最大的 200 个元素占据了信号总能量的 97.9%～99.7%。由此可见，若选择 Ω_1 作为稀疏表示矩阵，当 $n\geqslant 50$ 时，这些信号的能量都集中在了少量的大系数中，从而说明了稀疏表示矩阵 Ω_1 具有较强的稀疏信号能力。

3.2　稀疏表示基和观测矩阵的低相关性

因为确定一个感知矩阵是否满足有限等距特性是非常困难的，所以可以把稀疏表示基和观测矩阵之间的相关性大小作为感知矩阵是否满足有限等距特性的衡量标准[18]。相关性越小，有限等距特性越好。给定一组 N 维空间的正交基 (φ,ψ)，它们的相关性可以表示为：

$$\mu(\varphi,\psi)=\sqrt{N}\max|\langle\varphi_i^r,\psi_j\rangle|\in[1,\sqrt{N}] \tag{8}$$

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，我们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，通过计算它们之间的非相关性[18]来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定义为：

$$I(\varphi,\psi)=\min\|\theta_i\|_0,1\leqslant i\leqslant M \tag{9}$$

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

$$\theta_i=(\varphi^{\mathrm{T}}\varphi)^{-1}\varphi^{\mathrm{T}}\psi_i^r \tag{10}$$

其中，ψ_i^r 表示观测矩阵 ψ 的第 i 个行向量。$I(\varphi,\psi)$ 越大，φ 和

ψ 的非相关性越大，相关性越小。

表 2 列出了在每个数据集中，当稀疏表示基分别为 φ_1 和 φ_2 时 $I(\varphi,\psi)$的值。从表 2 中可以看出，无论选择哪一个稀疏表示基，$I(\varphi,\psi)$的值都较大，也就是说我们设计的稀疏表示基和观测矩阵之间的相关性较小，感知矩阵的有限等距性较好。选择 φ_1 和 φ_2 作为稀疏表示基时，$I(\varphi,\psi)$的值相差不大。结合 φ_1 和 φ_2 在图 1—图 3 中的表现可以发现，φ_1 的稀疏信号能力更强，效果更稳定，因此本文选择 φ_1 作为稀疏表示基，即 $\varphi=\varphi_1$。

表 2 稀疏表示基和观测矩阵的非相关性

Table 2 Non-correlation between sparse representation basis and measurement matrix

n	100	200	500	800	1000
KAIST-φ_1	98	187	465	776	975
KAIST-φ_2	99	189	483	780	988
MOTES-φ_1	99	189	467	776	976
MOTES-φ_2	99	189	486	782	990
GAS-φ_1	97	186	466	768	974
GAS-φ_2	99	188	484	782	987

4 同时稀疏贝叶斯学习

以往的多稀疏向量联合恢复算法需要提前已知多个稀疏向量的共同支持信息[19-23]，这在实际应用中较难实现，特别是在存在大量缺失数据的情况下。基于此，设计了一种基于稀疏贝叶斯学习的恢复算法，该算法可以通过学习获得部分支持信息，从而同时恢复多个稀疏向量。

首先，假设 $p(y_j|x_j)$满足方差为 σ^2 的高斯分布：$p(y_j|x_j;\sigma^2)=N(A_jx_j,\sigma^2\ I_{m_j})$。其中，$I_{m_j}$ 为一个辨识矩阵。令 $X=(x_1,x_2,x_3,\cdots,x_K)$，对于矩阵 X 的第 i 行，考虑以下两个高斯先验模型：

$$p(x_{ij}|M_i=1)\sim N(0,\gamma_i^b)(\text{行稀疏}) \tag{11}$$

$$p(x_{ij}|M_i=0)\sim N(0,\gamma_{ij}^s)(\text{元素稀疏}) \tag{12}$$

二进制序列 M_i 表示第 i 行的模型标签。γ_i^b 和 γ_{ij}^s 是未知的方差参数。行稀疏模型中该行所有元素的方差相同，而元素稀疏模型中每个元素都有其对应的方差。令 $M=\{M_1,M_2,M_3,\cdots,M_N\}$为所有行模型标签的集合。结合协同进化时间序列的特点可知，M 中至少有一部分为 1。每一个时间序列 s_j 定义一个对角矩阵$\Gamma_j\in R^{N\times N}$，它的第 i 个对角元素是 γ_i^b 还是 γ_{ij}^s 由该行的模型标签决定。令 $\theta=\{\sigma^2,M,\Gamma_1,\Gamma_2,\cdots,\Gamma_K\}$，$Y=\{y_1,y_2,\cdots,y_K\}$，本文的目标是找到使得概率 $p(Y;\theta)$最大的 θ。由于 Y 是已知的，因此一旦 θ 确定，我们就可以通过求解它的最大后验概率 $\arg\max\ p(X|Y;\theta)$来获得 X 的值。然而，直接计算 $p(Y;\theta)=\int p(Y|X;\theta)p(X;\theta)\mathrm{d}X$ 的最大值非常困难，因为要从 M 中挑选出最合适的模型需要解决 2^N 个不同的非凸优化问题。参考文献[24]中提出的方法，可利用变分贝叶斯理论将 $p(Y;\theta)$的计算转换成式(13)的形式：

$$\ln p(Y;\theta)=KL(q(X))\ \|\ p(X|Y;\theta))+F(q(X),\theta) \tag{13}$$

其中，变分分布 $q(X)$是后验概率分布 $p(X|Y;\theta)$的近似分布。因为 KL 散度是非负的，所以有 $\ln p(Y;\theta)\geqslant F(q(X),\theta)$，当且仅当 $q(X)=p(X|Y;\theta)$时等式成立。类似于 EM 算法，可以通过迭代 $q(X)$和 θ 来最大化 $\ln p(Y;\theta)$的值。

1)E 步：令 KL 散度为零，则变分分布 $q(x_j)$可以根据式(14)更新：

$$q(x_j)=p(x_j|y_j;\Gamma_j,\sigma^2,M)=N\sim(\mu_j,\Sigma_j) \tag{14}$$

其中，第二个等式的成立请参照文献[25]得出的结论；μ_j 和 Σ_j 分别按式(15)和式(16)计算：

$$\mu_j=\Gamma_j\ A_j^{\mathrm{T}}\ (\Sigma_j^y)^{-1}y_j \tag{15}$$

$$\Sigma_j=\Gamma_j-\Gamma_j\ A_j^{\mathrm{T}}\ (\Sigma_j^y)^{-1}A_j\ \Gamma_j \tag{16}$$

其中，$\Sigma_j^y=\sigma^2I+A_j\ \Gamma_j\ A_j^{\mathrm{T}}$。

2)M 步：将 $q(X)$代入 $F(q(X),\theta)$，则可得 θ 为：

$$\begin{aligned}\theta&=\arg\max\int q(X)\ln p(Y,X;\theta)\\&=\{\arg\max_{M,\Gamma_1,\cdots,\Gamma_K}\int q(X)\ln p(X;M,\Gamma_1,\cdots,\Gamma_K)\mathrm{d}X\\&=\arg\max_{\sigma^2}\int q(X)\ln p(Y|X;\sigma^2)\mathrm{d}X\}\end{aligned} \tag{17}$$

一旦 M 确定，那么参数 Γ_j 就可以更新为：

$$\{\Gamma_j^{opt}\}_{j=1}^K=\arg\max_{\{\Gamma_j\}_{i=1}^K}\int q(X)\ln p(X;M,\Gamma_1,\Gamma_2,\cdots,\Gamma_K)\mathrm{d}X \tag{18}$$

则有：

$$\gamma_i^b=\begin{cases}\dfrac{1}{K}\sum_{j=1}^K((\Sigma_j)_{i,i}+\mu_{i,j}^2), & M_i=1 \quad (19)\\ \gamma_{ij}^s=(\Sigma_j)_{i,i}+\mu_{i,j}^2, & M_i=0 \quad (20)\end{cases}$$

其中，$\mu_{i,j}$表示μ_j 的第 i 个元素，$(\Sigma_j)_{i,i}$表示 Σ_j 的第 i 个对角元素。然而，必须要找到 M 正确的更新方式，否则需要比较 2^N 次 $\int q(X)\ln p(X;M,\{\Gamma_j^{opt}\}_{j=1}^K)\mathrm{d}X$ 的值才能找到最好的 M。为了方便地推导出有效的模型选择方案，可以考虑 $\int q(X)\ln p(X;M,\{\Gamma_j^{opt}\}_{j=1}^K)\mathrm{d}X$ 的上界：

$$\int q(X)\ln p(X;M)\mathrm{d}X\leqslant\ln\int q(X)\ p(X;M)\mathrm{d}X \tag{21}$$

此处省略了参数$\{\Gamma_j^{opt}\}_{j=1}^K$，以保持式(21)的齐整。这个替代的目标函数使得我们可以独立地决定每行是行稀疏还是元素稀疏。将 X 视为“观测数据”，$p(X;M)$就成为了模型的证据。应用文献[26]提出的 BIC 理论，X 的第 i 行 x_i 的模型证据的近似值可计算为：

$$\begin{aligned}\ln p(x_i|M_i=1)&\approx\max\ln p(x_i|\gamma_i^b)-\frac{1}{2}\ln K+C_0\\&=-\frac{K}{2}\ln\frac{\sum_{j=1}^K x_{ij}^2}{K}-\frac{K}{2}(1+\ln 2\pi)-\\&\quad\frac{1}{2}\ln K+C_0\end{aligned} \tag{22}$$

$$\begin{aligned}\ln p(x_i|M_i=0)&\approx\max_{\gamma_{i1}^s,\cdots,\gamma_{iK}^s}\ln p(x_i|\gamma_{i1}^s,\cdots,\gamma_{iK}^s)-\\&\quad\frac{K}{2}\ln K+C_0\\&=-\frac{1}{2}\sum_{j=1}^K\ln x_{ij}^2-\frac{K}{2}(1+\ln 2\pi)-\\&\quad\frac{K}{2}\ln K+C_0\end{aligned} \tag{23}$$

其中，C_0 是一个常数。根据式(21)－式(23)，可以通过比较式(24)和式(25)来决定 M_i 的值：

$$\ln\int q(x_i)p(x_i;M_i=1)\mathrm{d}x_i=-\frac{K}{2}\ln\frac{2\pi\sum_{j=1}^{K}\mu_{i,j}^2+(\Sigma_j)_{i,i}}{K}-\frac{K}{2}-\frac{\ln K}{2}+C_0 \tag{24}$$

$$\ln\int q(x_i)p(x_i;M_i=0)\mathrm{d}x_i=-\frac{1}{2}\sum_{j=1}^{K}\ln(\mu_{i,j}^2+(\Sigma_j)_{i,i}-\frac{K}{2}(1+\ln 2K\pi)+C_0 \tag{25}$$

最后，可以对式(17)的第一项进行最大化来更新噪声方差：

$$(\sigma^2)^{\mathrm{new}}=\arg\max_{\sigma^2}\int q(X)\ln p(Y|X;\sigma^2)\mathrm{d}X=\frac{\sum_{j=1}^{K}\|y_j-A\mu_j\|_F^2+\sigma^2\sum_{j=1}^{K}\sum_{i=1}^{N}(1-(\Gamma_j)_{i,i}^{-1}(\Sigma_j)_{i,i})}{\sum_{j=1}^{K}m_j} \tag{26}$$

同时稀疏贝叶斯学习算法的步骤如下：

Step1　初始化 σ^2 和 $\Gamma_j=I$；

Step2　分别通过式(15)和式(16)计算 μ_j 和 Σ_j；

Step3　通过比较式(24)和式(25)更新 M_i；

Step4　根据式(19)和式(20)更新 Γ_j；

Step5　根据式(26)更新噪声方差 σ^2；

Step6　迭代 Step2－Step5 直到收敛。

5　仿真与分析

本节仍将使用上文描述的数据集进行仿真实验。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。

5.1　比较算法和性能评估标准

为了对所提算法(SSBL)的性能进行有效评估，将其与其他 3 种算法进行仿真比较：1)经典的基于样条插值算法(Spline Interpolation，SI)[4]；2)时域动态矩阵分解算法(TDMF)[10]；3)递归神经网络(RNN)方法[11]。

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，表示预测值与真实观测值之间的样本偏差。其定义如下：

$$RMSE=\frac{1}{N}\sqrt{\sum_{ij}(x_{ij}-\hat{x}_{ij})^2} \tag{27}$$

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，每种方法重复执行 50 次，计算其平均运行时间(Average Running Time，ART)。

$$ART=\frac{T}{50} \tag{28}$$

5.2　仿真结果与分析

图 4－图 6 分别给出了各种方法在小数据集 MOTES、中等数据集 KAIST 和大数据集 GSA 上的仿真结果。从图中可看出，就均方根误差而言，无论在哪个数据集上，所提算法相比其他算法都展示了更好的性能，从而说明该方法是适用并且非常有效的。具体以图 6 为例，当数据缺失率为 50%时，各算法的性能相差不大，但是随着数据缺失率增大，所提算法的优越性越来越明显，这充分说明即使在有大量数据缺失的情况下，本文算法依然能有效地恢复出整个时间序列。

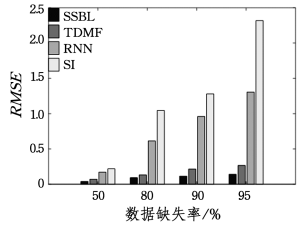

图 4　MOTES 中各算法的性能比较

Fig. 4　Performance comparison of proposed methods and other methods in MOTES

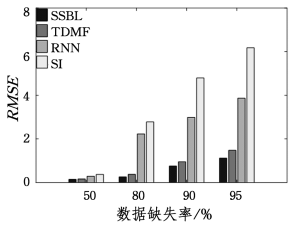

图 5　KAIST 中各算法的性能比较

Fig. 5　Performance comparison of proposed methods and others methods in KAIST

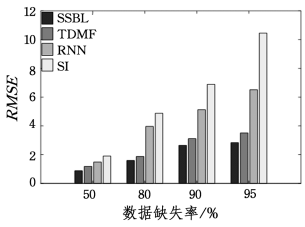

图 6　GSA 中各算法的性能比较

Fig. 6　Performance comparison of proposed methods and other methods in GSA

表 3 列出了当数据集维数一定时，所提算法与其他算法的平均运行时间。从表 3 中可以发现，当数据集较小时，样条插值算法(SI)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。随着数据集变大，本文方法(SSBL)的优越性开始凸显，而插值算法的运行时间明显变长，这是因为插值算法的运行时间是多个时间序列的叠加，而本文算法可以同时预测多个时间序列中的缺失值。稀疏向量的稀疏性对算法性能的影响如图 7 所示。由图 7 可知，算法的性能受数据集特性的影响，因此平均运行时间随维数的变化因数据集的不同而不同。仿真实验中 MOTES 的数据集维数为 8 * 1000，从表 3 中可以看出，就平均运行时间而言，所提算法在中等数据集 MOTES 上的性能与具有大规模平行处理能力的 RNN 算法相近；仿真实验中 GSA 的数据集维数为 16 * 1000，所提算法在该数据集上具有最短的平均运行时间。综上，在数据集维数较多且数据缺失严重的情况下，所提算法在均方根误差和平均运行时间上都具有优越的性能。

表 3　各算法的平均运行时间比较

Table 3　Comparison of average running time of proposed methods and others methods

(单位：s)

Data set	RNN	TDMF	SI	SSBL
KAIST	0.216	0.454	0.063	0.264
MOTES	0.368	0.609	0.168	0.436
GSA	0.756	1.487	0.672	0.741

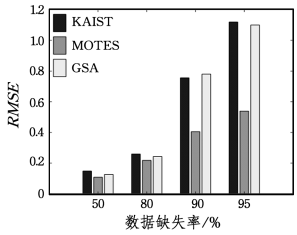

图 7 稀疏向量的稀疏性对算法性能的影响

Fig. 7 Effect of sparse vector on algorithm performance

为了研究稀疏向量的稀疏性对算法性能的影响，随机从 MOTES 数据集和 GSA 数据集中删除一些数据列，使得 3 个数据集的维数相同。由图 7 可以看出，所提算法在 MOTES 数据集上的表现优于其他数据集。回顾图 1－图 3 可以看出，所选择的稀疏矩阵在 MOTES 数据集下的稀疏向量具有最好的稀疏性，即当 n 相同时，$\sum_{i=1}^{n}(x_j^i)^2/\sum_{i=1}^{N}(x_j^i)^2$ 最大。因此可以得出结论，所提算法的性能与稀疏向量的稀疏性有关。稀疏向量越稀疏，均方根误差越小。

为了研究缺失数据类型对算法性能的影响，假设数据分别是随机、均匀或连续丢失的(见图 8－图 10)。当数据均匀缺失时，所提算法的性能最好。其原因是若数据是均匀缺失的，则能观测到的数据也是均匀的，我们可以获得每一段时间内的有用信息。但实际上，数据的缺失类型往往是随机的。我们可以观察到，在数据随机缺失的情况下，算法性能仍旧较好，甚至当数据连续缺失时，所提算法相比于其他算法产生的均方根误差依旧很小。这是因为本文算法充分利用了压缩感知原理，将缺失预测问题建模成了稀疏向量恢复问题，只要观测值的数量大于稀疏向量的稀疏度，就能基于这些观测值以高概率恢复出原始时间序列。

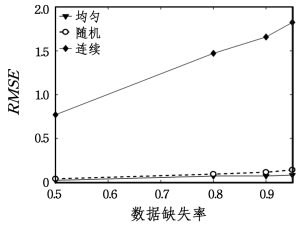

图 8 MOTES 中数据缺失类型对算法性能的影响

Fig. 8 Effect of type of data missing on algorithm performance in MOTES

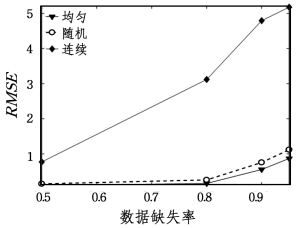

图 9 KAIST 中缺失类型对算法性能的影响

Fig. 9 Eeffect of type of data missing on algorithm performance in KAIST

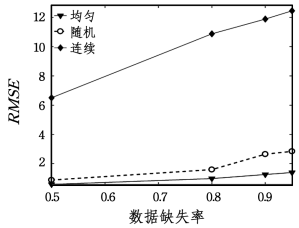

图 10 GSA 中缺失类型对算法性能的影响

Fig. 10 Effect of type of data missing on algorithm performance in GSA

结束语 本文提出了一种基于稀疏贝叶斯学习的缺失数据预测算法，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。此外，由于传统的观测矩阵不适用于本文的应用场景，因此设计了容易实现且与稀疏表示基相关性低的观测矩阵。通过分析和讨论，证明了可以利用压缩感知理论将协同进化时间序列中的缺失数据预测问题建模成多稀疏向量恢复问题。最后，针对协同进化时间序列的特点，本文设计了一种基于稀疏贝叶斯学习的恢复算法，该算法可以通过学习得到部分支持信息，从而同时解决多个稀疏向量的恢复问题。仿真结果表明，即使在数据大量缺失的情况下，所提算法也能有效地恢复出整个时间序列，尤其是在数据集较大、数据缺失比较严重的情况下具有明显的优势。本文设计的稀疏恢复算法是基于协同进化时间序列的特点设计的，当多个时间序列不是共进化时间序列时，解决稀疏恢复将花费较长的时间。下一步的研究方向是如何设计出一种普适性的多稀疏向量同时恢复算法，以同时预测多个时间序列中的缺失数据。

参 考 文 献

[1] SOWMYA R, SUNEETHA K R. Data mining with big data[C]// International Conference on Intelligent Systems and Control. IEEE, 2017: 246-250.

[2] SHI W, ZHU Y, YU P S, et al. Temporal dynamic matrix factorization for missing data prediction in large scale coevolving time series [J]. IEEE Access, 2017, 4(99): 6719-6732.

[3] ELENI I, VLAHOGIANNI J, GOLIAS C, et al. Short-term traffic forecasting: Overview of objectives and methods [J]. Transport Reviews, 2004, 24(5): 533-557.

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

[5] 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.

[6] AL-DEEK H M, CHANDRA C. New algorithms for filtering and imputation of real-time and archived dual-loop detector data in I-4 data warehouse[C] // Meeting of the Transportation-Research-Board. 2004: 116-126.

[7] BOYLES S D. A comparison of interpolation methods for missing traffic volume data[C] // Transportation Research Board Annual Meeting. 2011: 23-27.

[8] LIPPI M, BERTINI M, FRASCONI P. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(2): 872-882.

[9] LI Y, LI Z, LI L, et al. Comparison on PPCA, KPPCA and MPPCA based missing data imputing for traffic flow[C] // Proceedings of IEEE Conference on Intelligent Transportation System. 2013: 1535-1540.

[10] SHI W, ZHU Y, YU P. Effective Prediction of Missing Data on

Apache Spark over Multivariable Time Series[J]. IEEE Transactions on Big Data,2017,PP(99):1.

[11] STRAUMAN A S,BIANCHI F M,MIKALSEN K. Classification of postoperative surgical site infections from blood measurements with missing data using recurrent neural networks[C]// IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). 2018:307-310.

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

[13] HADI A,WAHIDAH I. Delay estimation using compressive sensing on WSN IEEE 802.15.4[C]// International Conference on Control, Electronics, Renewable Energy and Communications. IEEE,2017:192-197.

[14] ARJOUNE Y,KAABOUCH N,GHAZI H E,et al. Compressive sensing:Performance comparison of sparse recovery algorithms [C]// Computing and Communication Workshop and Conference. IEEE,2017:1-7.

[15] RHEE I,SHI N M,HONG S,et al. Mobility traces[OL]. http://carwdad.org/ncsu/mobilitymodels.

[16] SAMUEL M. Intel Lab Data[OL]. http://db.csail.mit.edu.

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

[18] 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. 2012:1-12.

[19] HU S W,LIN G X,HSIEH S H,et al. Performance analysis of joint-sparse recovery from multiple measurement vectors with prior information via convex optimization[C]// IEEE International Conference on Acoustics,Speech and Signal Processing. IEEE,2016:4368-4372.

[20] ZHAO X,YANG Q,ZHANG Y. Synthesis of sparse linear array with multiple patterns based on joint sparse recovery[C]// IEEE International Symposium on Antennas and Propagation & Usnc/ursi National Radio Science Meeting. IEEE, 2017: 425-426.

[21] WALEWSKI A C,STEFFENS C,PESAVENTO M. Off-Grid Parameter Estimation Based on Joint Sparse Regularization[C]// International Itg Conference on Systems,Communications and Coding. 2017.

[22] ZHANG Z,RAO B D. Sparse signal recovery in the presence of correlated multiple measurement vectors [C] // International Conference on Acoustics Speech and Signal Processing. 2010: 3986-3989.

[23] PRASAD R,MURPHY C R,RAO B D. Joint approximately sparse channel estimation and data detection in OFDM systems using sparse Bayesian learning[J]. IEEE Signal Processing Letters,2014,62(14):3591-3603.

[24] CHEN W. Simultaneous sparse Bayesian learning with partially shared support [J]. IEEE Signal Processing Letters, 2017, 24(11):1641-1645.

[25] TIPPING M E. Sparse Bayesian learning and the relevance vector machine[J]. Journal of Machine Learning Research,2001, 1(3):211-244.

[26] BISHOP C M. Pattern Recognition and Machine Learning (Information Science and Statistics)[M]. Springer-Verlag New York,2006.