计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700045

一种基于张量的车辆交通数据缺失估计新方法

张德干　范洪瑞　龚倡乐　高瑾馨　张　婷　赵彭真　陈　晨
天津理工大学计算机科学与工程学院计算机视觉与系统教育部重点实验室　天津 300384
天津理工大学计算机科学与工程学院智能计算及软件新技术天津市重点实验室　天津 300384

摘　要　面对当前庞大的智慧交通数据量,收集并统计处理是必要且重要的过程,但无法避免的数据缺失问题是目前的研究重点。文中针对车辆交通数据缺失问题提出一种基于张量的车辆交通数据缺失估计新方法:集成贝叶斯张量分解(Integrated Bayesian Tensor Decomposition,IBTD)。该算法在数据模型构建阶段,利用随机采样原理,将缺失数据随机抽取生成数据子集,并用优化后的贝叶斯张量分解算法进行插补。引入集成思想,将多个插补后的误差结果进行分析排序,考虑时空复杂度,择优平均得到最优结果。通过平均绝对百分比误差之后(Mean Absolute Percentage Error,MAPE)和均方根误差(Root Mean Square Error,RMSE)对提出模型的性能进行评估。实验结果表明,所提新方法能够有效地对不同缺失量的交通数据集进行插补,并能得到很好的插补结果。

关键词:交通数据;数据缺失;张量;随机采样;贝叶斯张量分解
中图法分类号　TP399

New Method of Data Missing Estimation for Vehicle Traffic Based on Tensor

ZHANG De-gan,FAN Hong-rui,GONG Chang-le,GAO Jin-xin,ZHANG Ting,ZHAO Peng-zhen and CHEN Chen
Key Laboratory of Computer Vision and System,Tianjin University of Technology,Tianjin 300384,China
Tianjin Key Lab of Intelligent Computing& Novel software Technology,Tianjin University of Technology,Tianjin 300384,China

Abstract　In the face of the current huge amount of intelligent traffic data,collecting and statistical processing is a necessary and important process,but the problem of inevitable data missing is the current research focus. Aiming at the problem of vehicle traffic data missing,this paper proposed a new method based on tensor for vehicle traffic data missing estimation,Integrated Bayesian tensor decomposition (IBTD). In the data model construction stage,the random sampling principle was used to randomly extract the missing data to generate a subset of data,and the optimized Bayesian tensor decomposition algorithm was used for interpolation. By introducing the integration idea,the error results after multiple interpolations were analyzed and sorted,consider the spatio-temporal complexity,and choose the optimal average to get the best result. The performance of the proposed model was evaluated by mean absolute percentage error (MAPE) and root mean square error (RMSE). Experimental results show that the proposed method can effectively interpolate the traffic datasets with different missing quantities and get good interpolation results.

Keywords　Traffic data,Data missing,Tensor,Random sampling,Bayesian tensor decomposition

1　引言

随着现代传感技术、通信技术、计算机技术与信息技术的快速发展,智能交通系统(Intelligent Transport System,ITS)被逐步推广应用于我国道路交通管理和控制,已成为提高交通安全和便捷性的有效手段[1-2]。交通信息采集系统是 ITS 的重要组成部分,通过获取全面、丰富、实时的交通信息可以把握城市道路交通状况与变化规律,为城市交通规划和决策提供科学依据。

在实际应用中所需数据应具有高空间和时间分辨率,才能达到建模、交通管理、预测和路线引导等目的,然而现实中往往出现大量的缺失数据和低质量数据[3-4]。缺失数据通常会产生非常广泛的影响,如果数据库中收集的是不完整的缺失数据,不仅会造成实际获取与预先估计数据量之间的差异,还会使最终计算的准确性降低。有些数据是不完整或存在缺失的,但系统将其看成是完整数据,就将形成数据处理误差。更有一些算法或者系统是在理想的非缺失数据集的基础上进行操作计算的,此时若发现数据集不完整,将造成计算过程直接停止[5-8]。

对于缺失数据的处理,国内外学者通常采取最典型的两

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(61571328);天津市重大科技专项(15ZXDSGX00050,16ZXFWGX00010);天津市科技支撑重点项目(17YFZCGX00360);天津市自然科学基金重点项目(18JCZDJC96800);天津市科技创新和 131 人才团队(TD12-5016,TD13-5025,No. 2015-23)
This work was supported by the National Natural Science Foundation of China (61571328),Major Project of Science and Technology in Tianjin (15ZXDSGX00050,16ZXFWGX00010),Tianjin Key Projects supported by Science and Technology (17YFZCGX00360),Tianjin Natural Science Foundation Key Project (18JCZDJC96800),Training Plan of Tianjin Science and Technology Innovation and 131 Talent Team (TD12-5016,TD13-5025,2015-23).
通信作者:高瑾馨(974281483@qq.com)

种方法:1)将数据集中有缺失的部分数据直接整段删除,只利用现有收集完整的整段数据用作交通流预测应用;2)采用算法将不完整数据补全成完整数据[9-11]。以上两种方法各有利弊,第一种方法毫无疑问是最直接有效的方法,但是不能充分利用所有的数据信息,特别是在所删除的数据发生的时空节点能体现重要数据信息的时候,将此类数据信息删除会大大降低预测交通流信息等应用的精确度[11-13]。相比之下,第二种方法逐渐得到该领域的广泛重视与研究,基于向量、矩阵与张量的数据修复方法陆续被提出,学者们从多个角度、多个方面针对所提方法做出优化和比较[14]。当存在严重的数据缺失时,此类方法往往能够稳定地表示出优于第一种方法的适用性,但同时也存在一些弊端,诸如修复时产生的数据误差也会降低整体性能[15-18]。针对上述问题,本文提出了一种基于张量的车辆交通数据缺失估计新方法——集成贝叶斯张量分解(IBTD)。

与传统缺失数据插补方法有所不同的是,本文提出的IBTD算法学习文献[19-26]理论基础,设置两个超参数并放置两个共轭先验,通过不断更新参数使模型快速收敛,在此基础上结合了能够更好表示数据时空关联性的张量模型并利用随机抽取生成数据子集并集成两大优势。本文提出的集成贝叶斯张量分解(IBTD)能够在基于高时空关联性的张量模型下,有效地修复缺失的交通数据,与传统方法相比表现出更好的插补性能。

本文选用的实验数据为交通流量数据,对该数据进行插补以对交通部署等工作提供科学数据。根据获取的数据,本文将历史数据库构建为 $13\times32\times180$ 大小的三维张量,可同时利用如空间模式、天模式等多种模式的时空相关信息,如图1所示,3个维度分别表示13个相关路段、32天、1天180个流量。

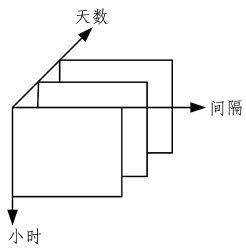


图1 交通数据3阶张量模型

Fig. 1 3rd order tensor model of traffic data

2 相关研究工作

目前,相关研究人员基于数据缺失问题已提出了多种具体算法。将缺失的数据利用已知数据的插值来填充,主要采用指数平滑和样条的方法。文献[20]建立并比对了很多回归模型的插补性能,得出二次回归模型的插补效果最为理想。文献[21]用贝叶斯主成分分析法(Bayesian Principal Componenit Analysis,BPCA)插补缺失的交通数据。文献[22]用张量分解的方法插补缺失数据。文献[23]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在数据量较小的情况下性能较差,而当数据量大时其计算代价又很大。文献[24]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表现出更好的预测性能。文献[25]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。文献[26]使用贝叶斯高思张量分解(Bayesian Gaussian CP,BGCP)插补缺失数据,将贝叶斯概率插补框架用于时空(或其他多维)设置中进行缺失数据插补。该方法通过在模型参数/超参数上放置灵活的先验/超前验,贝叶斯模型可以有效地自动表征数据的变化并避免过度拟合,但当数据集变大时,该分解算法可能需要大量冗余因子(r),特别是当不同维度的潜在因子的数量不平衡时,会影响算法性能并且计算代价非常昂贵。

本文提出的算法将从数据模型和集成优化两方面进行优化改进,使缺失数据插补结果更加理想。

3 模型构建

3.1 张量模型的基本思想

CP分解的主要思想是:一个高阶的张量可以看作是由若干个一维的因子矩阵构成的,那么就可以利用分解后的因子矩阵作计算。

本文考虑交通数据的特点,高阶张量虽然会一定程度上增加准确度,但与增加的复杂度相比,可行性不高,因此主要以三阶张量为例。

建立一个三阶张量 $Y\in\Re^{m\times n\times t}$,$m$ 为小时维数,n 为间隔维数,f 为天数维数。根据CP分解的基本思想,可构建因子矩阵 $\boldsymbol{A}\in\Re^{m\times r}$,$\boldsymbol{B}\in\Re^{n\times r}$,$\boldsymbol{X}\in\Re^{t\times r}$,$r$ 为张量 Y 的CP秩,由此可得到张量的低秩结构,如下所示:

$$Y=\sum_{s=1}^{r}a_{is}b_{js}x_{ts} \tag{1}$$

3.2 贝叶斯张量分解的基本原理

3.1节介绍的张量模型是缺失张量。本文用 Ω 表示那些观察到的元素的索引集,然后为数据生成过程引入完全贝叶斯模型。

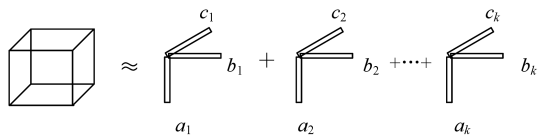


图2 三阶张量cp分解模型

Fig. 2 Third-order tensor cp decomposition model

首先,假设每个观察到的元素($i\in\Omega$)的噪声项都遵循独立的高斯分布:

$$y_{ijt}\sim\mathrm{N}(\sum_{s=1}^{r}a_{is}b_{js}x_{ts},\tau^{-1}) \tag{2}$$

其中,$\mathrm{N}(\cdot)$表示多元高斯分布;τ 是精度,它是所有元素的通用参数。

为了通过贝叶斯推断估计出因子矩阵,需要进一步设置共轭先验,此时共轭先验为多元高斯分布。由于高斯分布的两个参数都未知,为了对张量数据进行适当的建模,需在因子矩阵组和精度 τ 上放置敏捷的先验分布。对于因子矩阵,假设其行向量的先验分布为多元高斯,具体表现形式如式(3)所示:

$$\begin{aligned}a_i&\sim\mathrm{N}(\eta_a^l,(\Sigma_a^l)^{-1})\\b_j&\sim\mathrm{N}(\eta_b^l,(\Sigma_b^l)^{-1})\\x_t&\sim\mathrm{N}(\eta_x^l,(\Sigma_x^l)^{-1})\end{aligned} \tag{3}$$

其中，a_i，b_j，x_t 为因子矩阵 $\boldsymbol{A}$，$\boldsymbol{B}$，$\boldsymbol{X}$ 的第 i 行，第 j 行和第 t 行。

为了增强模型的稳健性，与传统贝叶斯设置不同，本文放置两个共轭先验超参数 $\eta^l \in \Re^r$ 和 $\Sigma^l \in \Re^{r\times r}$，相关基础理论可参见文献[26]。

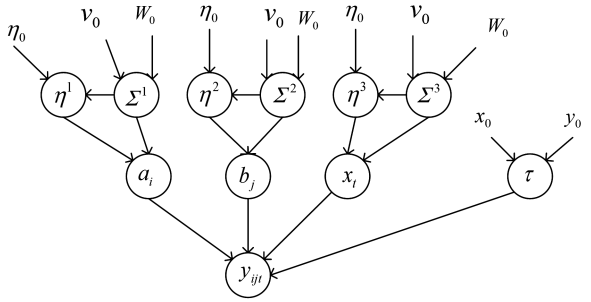


图 3 贝叶斯张量分解概率图模型

Fig. 3 Bayesian tensor decomposition probability map model

在式(2)中的高斯假设下，精度参数 τ 捕获数据中的噪声程度，τ 服从伽马分布：

$$\tau \sim Gamma(x_0, y_0) \tag{4}$$

其中，x_0 和 y_0 分别是形状参数和尺度参数。

基于上述假设，可得到一个贝叶斯网络。对于模型参数 a_i，b_j，x_t，其先验是一个多元高斯分布，则其后验分布也是一个多元高斯分布，如式(5)所示：

$$\begin{aligned} a_i &\sim \mathrm{N}(\overset{\wedge}{\eta}_a^l, (\overset{\wedge}{\Sigma}_a^l)^{-1}) \\ b_j &\sim \mathrm{N}(\overset{\wedge}{\eta}_b^l, (\overset{\wedge}{\Sigma}_b^l)^{-1}) \\ x_t &\sim \mathrm{N}(\overset{\wedge}{\eta}_x^l, (\overset{\wedge}{\Sigma}_x^l)^{-1}) \end{aligned} \tag{5}$$

其中，后验参数更新式为：

$$\begin{aligned} \overset{\wedge}{\Sigma}_a &= \tau \sum_{j,t_1:(i,j,t)\in\Omega} (b_j \otimes x_t)(b_j \otimes x_t)^{\mathrm{T}} + \Sigma \\ \overset{\wedge}{\Sigma}_b &= \tau \sum_{i,t_1:(i,j,t)\in\Omega} (a_i \otimes x_t)(a_i \otimes x_t)^{\mathrm{T}} + \Sigma \\ \overset{\wedge}{\Sigma}_x &= \tau \sum_{i,j_1:(i,j,t)\in\Omega} (a_i \otimes b_j)(a_i \otimes b_j)^{\mathrm{T}} + \Sigma \end{aligned} \tag{6}$$

$$\begin{aligned} \overset{\wedge}{\eta}_a &= \tau \overset{\wedge}{\Sigma}_a^{-1} \sum_{j,t_1:(i,j,t)\in\Omega} (b_j \otimes x_t) y_{ijt} \\ \overset{\wedge}{\eta}_b &= \tau \overset{\wedge}{\Sigma}_b^{-1} \sum_{i,t_1:(i,j,t)\in\Omega} (a_i \otimes x_t) y_{ijt} \\ \overset{\wedge}{\eta}_x &= \tau \overset{\wedge}{\Sigma}_x^{-1} \sum_{i,j_1:(i,j,t)\in\Omega} (a_i \otimes b_j) y_{ijt} \end{aligned} \tag{7}$$

其中，$\otimes$ 表示点乘；Ω 表示张量 y 中所有被观测到的元素索引集合；$j,t(i,j,t)\in\Omega$ 表示切片 $y_{:jt}$ 中被观测到的元素索引；$i,t(i,j,t)\in\Omega$ 表示切片 $y_{:it}$ 中被观测到的元素索引；$i,j(i,j,t)\in\Omega$ 表示切片 $y_{:ij}$ 中被观测到的元素索引。

τ 的后验分布，也是由 x_0 和 y_0 参数化的 Gamma 分布：

$$\begin{aligned} P(\tau_\varepsilon \mid T, a^1, a^2, a^3, x_0, y_0) &\propto L(T \mid a^1, a^2, a^3, \tau_\varepsilon) \times Gamma(\tau_\varepsilon \mid x_0, y_0) \\ &\propto Gamma(\tau_\varepsilon \mid \hat{x}_0, \hat{y}_0) \end{aligned} \tag{8}$$

其中：

$$\begin{aligned} \hat{x}_0 &= \frac{1}{2} \sum_{i,j,k\in\Omega} 1 + x_0 \\ \hat{y}_0 &= \frac{1}{2} \sum_{i\in\Omega} (T_i - \hat{T}_i)^2 + y_0 \end{aligned} \tag{9}$$

3.3 采样新策略

集成学习实际就是通过建立几个模型组合来解决单一预测问题。它的工作原理是多个弱学习器相互独立的训练学习并作出预测判断。多个预测结果最终组合形成单预测，这种组合形成单预测的方法得到的结果优于其中任何一个单学习器的预测结果[27]。

在集成算法中，装包(bagging)方法就是在初始数据训练集的随机生成数据子集上通过多个同类或不同类的黑盒估计器训练数据，然后将所有弱学习器的预测结果通过一定的数据处理得到最终的预测结果。该方法在数据模型的构建中采用随机抽取生成随机数据子集的手段降低弱学习器的数据预测方差。在多数情况下，bagging 方法提供了一种非常简单的方式来改进单一模型，而无需修改背后的算法[28]。因为 bagging 方法可以减小过拟合，所以通常在强分类器和复杂模型上使用时表现的很好(例如完全决策树(fully developed decision trees))。本文算法主要利用 bagging 的随机抽取思想，得到多个随机数据子集，用于集成得到最优结果。

引理 1 Bagging 策略中弱学习器之间相互独立。

证明：随机采样(bootsrap)就是从初始数据训练集中有放回地抽取一定数量的数据生成随机数据子集。也就是说，已经被抽取过的数据在放回后还有可能继续被抽取。对于 Bagging 算法，通常选择随机抽取和训练集样本数 m 相同的样本。如此随机抽取后得到的随机数据子集和初始训练集样本的个数一致，但是每个数据集的内容都不同。如果对有 m 个样本训练集进行 T 次随机采样，则由于随机性，T 个采样集各不相同，也就是弱学习器之间相互独立。

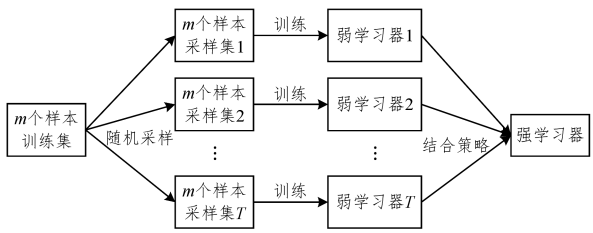


图 4 Bagging 随机采样原理图

Fig. 4 Schematic diagram of Bagging random sampling

3.4 择优排序机制

由于本文提出的 IBTD 旨在找出最优的插补结果，因此在得到集成插补的结果后，需对所有结果进行数据处理，按插补误差将数据排序，最后取若干个数据插补结果，将缺失部分的插补数据做平均处理，得到最后的插补结果以提升插补性能。

为了综合时耗与性能，本文分别集成 10，20，…，100 次，计算 IBTD 时间消耗和对应误差性能，得到如图 5 所示的集成次数与消耗时间和 RMSE 误差的关系。通过两条曲线的走向趋势可以看出，随着集成次数的增加，时间消耗也不断增加，但 RMSE 后期的变化并不明显，这说明集成次数的增加，会使误差降低，但考虑到时耗问题，性价比不高。因此，本文采取 40 次作为本组实验的集成次数。

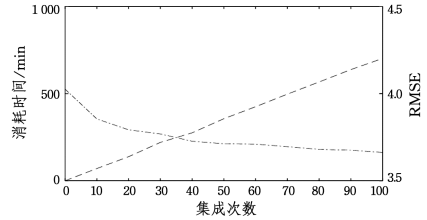


图 5 集成次数与消耗时间和 RMSE 误差关系

Fig. 5 Relationship between number of integrations, consumption timeand RMSE error

确定了集成次数后，将 40 个集成结果的误差结果按从小到大的顺序排序，分别取前 5，10，…，40 个结果进行结果平均。以取前 5 个结果为例，将误差最小的 5 个修复后的数据取出，针对每一个修复后的数据取 5 组的平均值，得到一个新的修复后的完整数据，再与原始数据对比得出误差数据。最后得到如图 6 所示的择优数量与 RMSE 的关系图。

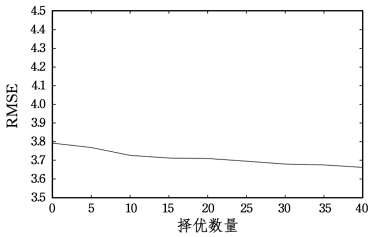


图 6 择优数量与 RMSE 的关系

Fig. 6 Relationship between preferred quantity and RMSE

由图 6 可知，随着抽取结果数量的增加，RMSE 的值不断降低，但抽取数为 10 以上对降低趋势趋于平缓。综合考虑复杂度问题，本文选择抽取前 10 个最优结果进行择优平均。

对比已有的多种排序机制的性能进行比较分析[29-30]后，本文将采用冒泡排序机制来对插补结果进行排序。

4 算法设计与分析

4.1 基于张量的车辆交通数据缺失估计算法

基于上文设计的模型、策略和机制，本文设计如下的基于张量的车辆交通数据缺失估计算法——集成贝叶斯张量分解算法(IBTD)，该算法的主体步骤如下。

1)将交通数据按路段 * 天 * 日流量的形式生成三阶张量数据模型。生成张量模型时，以随机性缺失和非随机性缺失两种情况，建立不同的缺失数据模型，用于评估算法性能。

2)对生成好的缺失张量数据用随机采样算法得到与原始缺失张量数据不同的不完整随机张量数据集。随机采样算法基于引理 1 Bagging 策略中弱学习器之间相互独立的性质，有放回地抽取数据，并生成数据集用于后续模型训练。此处调用算法 1。

3)将生成后的不完整随机张量数据集通过贝叶斯张量分解算法进行插补，通过放置的灵活先验参数，根据引理 1 后验概率正比于先验概率和似然度的乘积的原理，可以由先验和似然函数推出后验分布，然后不断更新超参数，直至收敛。利用集成思想循环插补，由于每一次插补的初始数据集不同，也将会得到各不相同的插补结果。此处调用算法 2。

4)将所有插补结果的误差参数进行冒泡排序、择优，将择优后的插补数据进行算数平均处理，得到更贴近原始数据的修复数据。此处调用算法 3。

算法 1 采样新策略算法

输入：样本 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，基修复器贝叶斯张量算法，基修复器迭代次数 t

输出：最终的修复器 f(x)

1. 对于 q=1,2,…,t

1.1 对训练集进行第 t 次随机采样，共采集 m 次，得到包含 m 个样本的采样集 T_t

1.2 用采样集 T_t 训练第 t 个弱学习器 $G_t(x)$

2. 对 Q 个基修复器得到的插补结果进行择优算术平均得到的值为最终的模型输出。

算法 2 贝叶斯张量分解算法

计算共轭先验超参数 $\hat{W}^l,\hat{v}^l,\hat{\eta}^l,\hat{\Sigma}^l$，得到先验分布；

根据式(6)、式(7)，计算 $\hat{\eta}_a,\hat{\eta}_b,\hat{\eta}_x,\hat{\Sigma}_a,\hat{\Sigma}_b,\hat{\Sigma}_x$，得到后验分布；

根据式(9)，计算 $\hat{x}_0,\hat{y}_0$，得到精度参数的 Gamma 分布；

重复更新参数直至收敛。

For l=1,…,s

Compute $\hat{W}^l$、$\hat{v}^l$、$\hat{\eta}^l$、$\hat{\Sigma}^l$

Sample $\Sigma^l\sim W(\hat{W}^l,\hat{v}^l)$

Compute $\hat{\eta}^l$ and $\hat{\Sigma}^l$

Sample $\eta^l\sim N(\hat{\eta}^l,(\hat{\Sigma}^l)^{-1})$

$a_i\sim N(\eta_a^l,(\Sigma_a^l)^{-1})$

Sample $b_j\sim N(\eta_b^l,(\Sigma_b^l)^{-1})$ // get the prior distribution

$x_t\sim N(\eta_x^l,(\Sigma_x^l)^{-1})$

For k=1,…,l

Compute $\hat{\eta}_a,\hat{\eta}_b,\hat{\eta}_x$ and $\hat{\Sigma}_a,\hat{\Sigma}_b,\hat{\Sigma}_x$

$a_i\sim N(\hat{\eta}_a^l,(\hat{\Sigma}_a^l)^{-1})$

Sample $b_j\sim N(\hat{\eta}_b^l,(\hat{\Sigma}_b^l)^{-1})$// Update parameter

$x_t\sim N(\hat{\eta}_x^l,(\hat{\Sigma}_x^l)^{-1})$

End for

End for

Compute $\hat{x}_0$ and $\hat{y}_0$

Sample $\tau_\varepsilon\sim$Gamma($\hat{x}_0,\hat{y}_0$) / get the Gamma distribution

Until maximum number of iterations

算法 3 冒泡排序择优机制算法

算法描述：1)比较相邻的元素，如果第一个比第二个大，则交换两个位置；2)对每一对相邻元素作同样的工作；3)针对所有的元素重复以上步骤，除最后一个。

For i=1 to length[A]

For j=length[A] to i+1 // set collection of error results to A

If A[j] < A[j−1]

exchange A[j] and A[j−1]

End if

End for

End for //get sorted data order

Result=(A[1]+A[2]+…+A[10])/10 // Arithmetic result

4.2 IBTD 算法的伪代码描述

IBTD 算法伪代码如算法 4 所示。

算法 4

Input：a k-th order tensor T and the corresponding indicator tensor P

Initialization：we set t=40，$\eta_0=0$，$\beta_0=1$，$W_0=I$ (identity matrix)，$v_0=r$(the low rank)，$x_0=1$，and$y_0=1$

Begin

For q=1,…t do //Calling algorithm 2

$T_t=\Im(T,T_{bs})$ // we set the base fixer to?

For l=1,…,s

Compute $\hat{W}^l$、$\hat{v}^l$、$\hat{\eta}^l$、$\hat{\Sigma}^l$

Sample $\Sigma^l\sim W(\hat{W}^l,\hat{v}^l)$

Compute $\hat{\eta}^l$ and $\hat{\Sigma}^l$

Sample $\eta^l\sim N(\hat{\eta}^l,(\hat{\Sigma}^l)^{-1})$

$a_i\sim N(\eta_a^l,(\Sigma_a^l)^{-1})$

Sample $b_j \sim N(\eta_b^l, (\Sigma_b^l)^{-1})$ // get the prior distribution

$x_t \sim N(\eta_x^l, (\Sigma_x^l)^{-1})$

For k=1,…,l

Compute $\hat{\eta}_a, \hat{\eta}_b, \hat{\eta}_x$ and $\hat{\Sigma}_a, \hat{\Sigma}_b, \hat{\Sigma}_x$

$a_i \sim N(\hat{\eta}_a^l, (\hat{\Sigma}_a^l)^{-1})$

Sample $b_j \sim N(\hat{\eta}_b^l, (\hat{\Sigma}_b^l)^{-1})$ // Update parameter

$x_t \sim N(\hat{\eta}_x^l, (\hat{\Sigma}_x^l)^{-1})$

End forEnd for

Compute $\hat{x}_0$ and $\hat{y}_0$

Sample $\tau_\varepsilon \sim$ Gamma($\hat{x}_0, \hat{y}_0$) // get the Gamma distribution

Until maximum number of iterations

End for

For i=1 to length[A]

For j=length[A] to i+1 // set collection of error results to A

If A[j] < A[j-1]

exchange A[j] and A[j-1]

End if

End for

End for //get sorted data order

Result=(A[1]+A[2]+…+A[10])/10 // Arithmetic result

End

5 实验测试与对比分析

5.1 速度数据

本文将长沙市芙蓉区局部路网的交通流数据作为研究对象，如图 7 所示，路网中包括 4 个交叉口（编号为 12001，11605，12701 和 12700），其中待测路段（编号 8）长约 400 m，位于嘉雨路与万家丽中路中间的远大一路由东往西方向上的一段。待测路段的二阶上游包括路段 1－6，二阶下游包括路段 7，9，10－13。（数据来源：Openits）

图 7 长沙市芙蓉区局部路网图

Fig. 7 Diagram of local road network in Furong District, Changsha

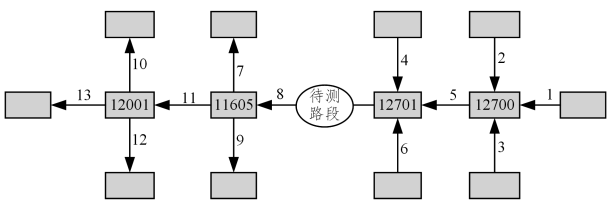


图 8 长沙市芙蓉区局部路网结构图

Fig. 8 Structure of local road network in Furong District, Changsha

5.2 数据张量结构表示

本文采用的数据集为长沙市交通流量数据，来源于目标区域各交叉口每个车道的环形线圈的实时采集，假定路段无出入口，交叉口进口道的流量即为对应路段的流量。流量数据由 SCATS Traffic Reporter 系统输出，各检测器每天收集 180 个流量数据。数据库包含了这 13 个路段 2013 年 9 月 17 日－10 月 18 日共 32 天的数据。

图 9 反映了不同路段在 9 月 22 日的交通流量变化趋势，虽然不同路段在流量大小上存在差异，但整体流量变化趋势具有很强的相似性，这与交通流在各个路段的空间转移有关。图 10 则反映了路段 1 在不同日期的流量变化趋势，由于交通流量与出行规律间存在必然联系，且人们的出行存在一定的周期性，除重大节假日等特殊情况下，每天的出行时间都存在早晚高峰，因此交通流波动相似。

通过构建路段 * 天 * 日流量三维张量，充分利用了数据的各种时空相关信息，因此可以更好地进行交通流数据修复。

本文通过随机删除一定数量的条目来创建数据集，从而将原始数据分成两组：观察到的（Ω）和缺失的（移除的）。对于这些“缺失”的条目，文中也有相应的基本事实，这使我们可以通过输入删除的条目直接评估模型的插补性能。

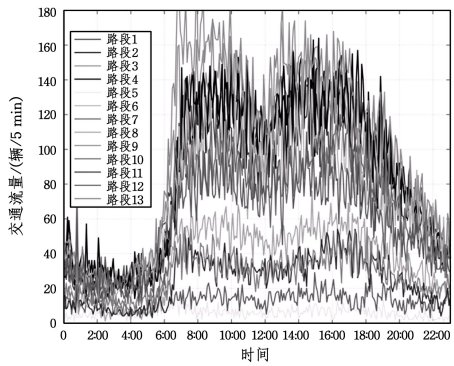


图 9 不同路段 9 月 22 日流量趋势图

Fig. 9 Flow trend chart of different road sections on September 22

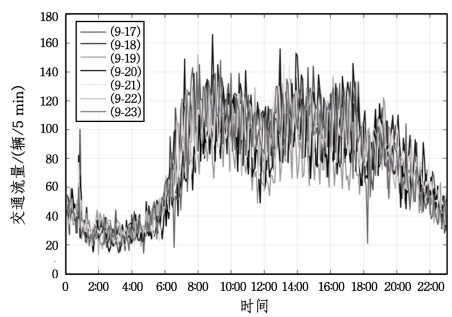
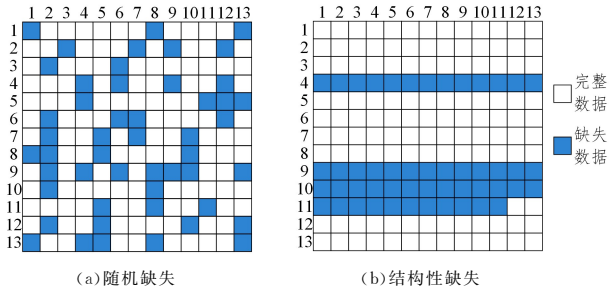


图 10 路段 1 不同日期流量趋势图

Fig. 10 Flow trend chart of road segment 1 on different dates

5.3 数据缺失情况

本实验评估了不同模型和不同数据在不同丢失率以及不同的丢失情景（随机缺失与非随机缺失）下的表现。随机缺失即在 13 个路段 32 天的历史数据中按照 10％的间隔随机抽取 10％，20％，…，90％的数据，将缺失数据设为 0。结构性缺失在同样的历史数据中抽取同样数量的数据假设缺失，但它假设数据缺失发生在同一时段不同路段上。以缺失 30％的数据为例，图 11 为随机缺失与结构性缺失两种数据缺失形式。



(a)随机缺失 (b)结构性缺失

图 11 数据缺失类型

Fig. 11 Data missing type

5.4 数据修复结果及分析

本文采用两种误差来衡量缺失数据的修复效果，即平均绝对百分比误差 MAPE 和均方根误差 RMSE。误差的计算公式如下：

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\frac{|T_i-\hat{T}_i|}{T_i}$$
$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{T}_i-T_i)^2} \tag{10}$$

在数值实验中，本文将 IBTD 模型与另外 4 种基于张量分解的插补方法进行比较：贝叶斯高斯张量分解(BGCP)、经典的张量 cp 分解、cp-wopt 算法、贝叶斯主成份分析。

根据式(10)计算得到的 RMSE 和 MAPE 误差结果如图 12、图 13 所示，可以看出在数据缺失类型为随机缺失的条件时，本文提出的 IBTD 算法能够稳定地降低修复误差，而且误差相对较小。随着缺失百分比的增大，5 种算法的误差都增大，但 cp 分解的增长幅度最大，说明 cp 分解算法对于缺失数据量更敏感，插补稳定性差。cp-wopt 插补性能优于传统 cp 分解，BPCA 算法性能稍差于 cp-wopt 算法。贝叶斯高斯张量分解与 IBTD 的增幅趋势大致相同，但相比之下 IBTD 的误差更小，插补效果更为理想。

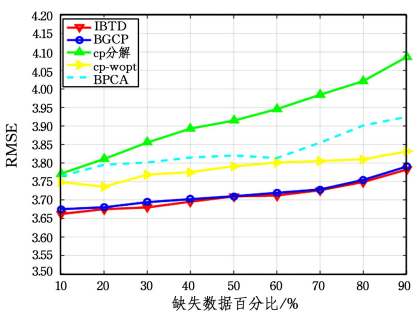


图 12 随机缺失条件下的数据修复均方根误差

Fig. 12 RMSE of data repair under random missing conditions

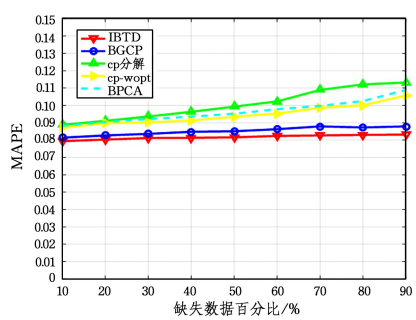


图 13 随机缺失条件下的数据修复平均绝对百分比误差

Fig. 13 MAPE of data repair under random missing conditions

由图 14 可看出，在数据缺失类型为结构性缺失的条件下，5 种算法的整体误差率都增大，说明结构性缺失比随机缺失对数据插补的影响更大。5 种算法相比之下，本文提出的 IBTD 在缺失数据百分比增大时，表现出更稳定的误差率，体现了本文算法使用集成择优原理的优越性。

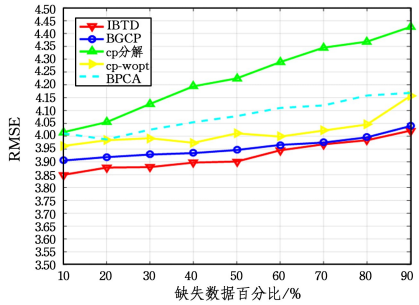


图 14 结构性缺失条件下的数据修复均方根误差

Fig. 14 RMSE of data repair under structural missing conditions

结束语 本文针对智能交通数据缺失问题，提出了一种基于张量的车辆交通数据缺失估计新方法——集成贝叶斯张量分解。对于不完整数据集，鉴于城市交通数据中强大的时空相关性，本文建立了张量模型，并表现出很好的性能。在贝叶斯张量分解算法的基础上加入随机采样和集成思想，通过随机抽取再修复得到多个插补结果，并取优平均，以降低插补误差，更精准地修复缺失数据，修复后的交通数据可以作为城市交通规划的科学判断依据。本文所提算法对于其他同类型插补算法有更稳定、更精确的插补效果，但还存在复杂度的问题。下一步研究可以针对降低本算法复杂度方向上进行改进，在保证修复准确性的同时降低时耗等复杂度问题。

参 考 文 献

[1] TANG Y M. Novel Reliable Routing Method for Engineering of Internet of Vehicles Based on Graph Theory[J]. Engineering Computations, 2019, 36(1): 226-247.

[2] GOULART J. Traffic data imputation via tensor completion based on soft thresholding of tucker core[J]. Transportation Research Part C: Emerging Technologies, 2017, 85(11): 348-362.

[3] TAN H, WU Y, SHEN B. Short-term traffic prediction based on dynamic tensor completion[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(8): 1524-1530.

[4] ZHANG T, ZHANG J. A Kind of Effective Data Aggregating Method Based on Compressive Sensing for Wireless Sensor Network[J]. EURASIP Journal on Wireless Communications and Networking, 2018, 2018(159): 1-15.

[5] ZHANG T. Novel Optimized Link State Routing Protocal Based on Quantum Genetic Strategy for Mobile Learning[J]. Journal of Network and Computer Applications, 2018, 2018(122): 37-49.

[6] LIU S. Novel Unequal Clustering Routing Protocol Considering Energy Balancing Based on Network Partiton & Distance for Mobile Education[J]. Journal of Network and Computer Applications, 2017, 88(15): 1-9.

[7] ZHOU S. A low duty cycle efficient MAC protocol based on self-adaption and predictive strategy[J]. Mobile Networks & Applications, 2018, 23(4): 828-839.

[8] CHEN C. New Method of Energy Efficient Subcarrier Alloca-

tion Based on Evolutionary Game Theory[J]. Mobile Networks and Applications,2019,24(10):30-45.
[9] GAO J X. Novel Approach of Distributed&Adaptive Trust Metrics for MANET[J]. Wireless Networks,2019,25(6):3587-3603.
[10] ZHANG D G,GE H. New Multi-hop Clustering Algorithm for Vehicular Ad Hoc Net works[J]. IEEE Transactions Intelligent Transportation Systems,2019,20(4):1517-1530.
[11] LIU S. Dynamic Analysis For The Average Shortest Path Length of Mobile Ad Hoc Networks under Random Failure Scenarios[J]. IEEE Access,2019,7:21343-21358.
[12] DUAN Y,LV Y,LIU Y L. An efficient realization of deep learning for traffic data imputation[J]. Transportation Research Part C:Emerging Technologies,2016,72(11):168-181.
[13] TANG J,ZHANG G,WANG Y. A hybrid approach to integrate fuzzy c-means based imputation method with genetic algorithm for missing traffic volume data estimation[J]. Transportation Research Part C:Emerging Technologies,2015,51(2):29-40.
[14] SUN L. Understanding urban mobility patterns with a probabilistic tensor factorization framework[J]. Transportation Research Part B:Methodological,2016,91(9):511-524.
[15] LI X Y,YUAN J Q. Empirical likelihood method for quantities with response data missing at random[J]. Acta. Mathematicae Applicatae Sinica,2012,28(2):265-274.
[16] NIU H L. Novel PEECR-based Clustering Routing Approach [J]. Soft Computing,2017,21(24):7313-7323.
[17] CHEN Y L. Simultaneous tensor decomposition and completion using factor priors[J]. IEEE Trans. Pattern Anal. Mach. Intell,2014,36(3):577-591.
[18] BAE B. Missing data imputation for traffic flow speed using spatio-temporal cokriging[J]. Transportation Research Part C:Emerging Technologies,2018,88(3):124-139.
[19] ZHAO Q. Bayesian CP factorization of incomplete tensors with automatic rank determination[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(9):1751-1763.
[20] ALDEEK H M. New algorithms for filtering and inputation of real-time and archived dual-loop detector data in I-4 data warehouse[C]// Meeting of the Transpoetation-Research-Board. 2014:116-126.
[21] QU L,LI L,ZHANG Y. PPCA-based missing data imputation for traffic flow volume: A systematical approach[J]. IEEE Trans. Intell. Transp. Syst,2009,10(3):512-522.
[22] TANETA H. A tensor-based method for missing traffic data completion[J]. Transp. Res. C,Emerg. Technol,2013,28(3):15-27.
[23] CAI Y. Fast mining of a network of coevolving time series[C]// 2015 SIAM International Conference on Data Mining. 2015:298-306.
[24] LI Y. Comparison on PPCA,KPPCA and MPPCA based missing data imputing for traffic flow[C]// IEEE Conference on Intelligent Transportation System. 2013:1535-1540.
[25] STRAUMAN A S. Classification of postoperative surgical site infections from blood measurement with missing data using recurrent neural networks[C]// IEEE EMBS International Conference on Biomedical & Health Informatics,2018:307-310.
[26] CHEN X Y. A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation[J]. Transportation Research Part C Emerging Technologies,2018,11(1):73-84.
[27] ZHANG Y N. An Improved Random Sampling Approach for Large Data Set Mining[C]// International Conference on Smart City and Systems Engineering (ICSCSE). 2016:25-26.
[28] YOSHIKI K. A Parallel Sampling Method for Bayesian Networks[C]// International Symposium on Computer Science and Intelligent Controls (ISCSIC). 2018:1-10.
[29] EDJLAL R. A sort implementation comparing with Bubble sort and Selection sort[C]// International Conference on Computer Research & Development. IEEE,2011,1(1):1-9.
[30] ARORA N. A Novel Sorting Algorithm and Comparison with Bubble sort and Insertion sort[J]. International Journal of Computer Applications,2012,23(23):91-100.

ZHANG De-gan, born in 1969, Ph. D, professor, is a member of IEEE in 2001. His research interest includes ITS, WSN, IOT, etc.

GAO Jin-xin, born in 1994, Ph.D candidate, is a member of IEEE in 2016. Her research interest includes WSN, industrial application, etc.