

# 顺序敏感的多源感知数据填补技术\*

马茜, 谷峪, 李芳芳, 于戈



(东北大学 计算机科学与工程学院, 辽宁 沈阳 110819)

通讯作者: 谷峪, E-mail: guyu@mail.neu.edu.cn

**摘要:** 近年来,随着感知网络的广泛应用,感知数据呈爆炸式增长,但是由于受到硬件设备的固有限制、部署环境的随机性以及数据处理过程中的人为失误等多方面因素的影响,感知数据中通常包含大量的缺失值,而大多数现有的上层应用分析工具无法处理包含缺失值的数据集,因此对缺失数据进行填补是不可或缺的.目前也有很多缺失数据填补算法,但在缺失数据较为密集的情况下,已有算法的填补准确性很难保证,同时未考虑填补顺序对填补精度的影响.基于此,提出了一种面向多源感知数据且顺序敏感的缺失值填补框架 OMSMVI(order-sensitive missing value imputation framework for multi-source sensory data).该框架充分利用感知数据特有的多维度相关性:时间相关性、空间相关性、属性相关性,对不同数据源间的相似度进行衡量;进而,基于多维度相似性构建以缺失数据源为中心的相似图,并将已填补的缺失值作为观测值用于后续填补过程中.同时考虑缺失数据源的整体分布,提出对缺失值进行顺序敏感的填补,即:首先对缺失值的填补顺序进行决策,再对缺失值进行填补.对缺失值进行顺序填补能够有效缓解在缺失数据较为密集的情况下,由于缺失数据源的完整近邻与其相似度较低引起的填补精度下降问题;最后,对 KNN 填补算法进行改进,提出一种新的基于近邻节点的缺失值填补算法 NI(neighborhood-based imputation),该算法利用感知数据的多维度相似性对缺失数据源的所有近邻节点进行查找,解决了 KNN 填补算法  $K$  值难以确定的问题,也进一步提高了填补准确性.利用两个真实数据集,并与基本填补算法进行对比,验证了算法的准确性及有效性.

**关键词:** 缺失数据;密集缺失;感知网络;顺序敏感的填补;多维度相关性

**中图法分类号:** TP311

中文引用格式: 马茜, 谷峪, 李芳芳, 于戈. 顺序敏感的多源感知数据填补技术. 软件学报, 2016, 27(9): 2332-2347. <http://www.jos.org.cn/1000-9825/5045.htm>

英文引用格式: Ma Q, Gu Y, Li FF, Yu G. Order-Sensitive multi-source sensory missing value imputation technology. Ruan Jian Xue Bao/Journal of Software, 2016, 27(9): 2332-2347 (in Chinese). <http://www.jos.org.cn/1000-9825/5045.htm>

## Order-Sensitive Missing Value Imputation Technology for Multi-Source Sensory Data

MA Qian, GU Yu, LI Fang-Fang, YU Ge

(School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)

**Abstract:** In recent years, it is recognized that sensing data is growing explosively with widespread use of sensing network. Due to the inherent hardware limitation, the randomness of distribution environment and unconscious errors during data processing, a deluge of missing values are mingled in original sensing data. Thus, imputing the missing values is essential because most of the existed analysis tools are not competent to the data sets containing missing values. So far, there have been many missing data imputation algorithms, however the accuracy of these algorithms is difficult to be guaranteed in the scenario of lumped missing data. Besides, these existing algorithms don't take the imputation order which influences the imputation accuracy into consideration. To address the above issues, this paper proposes an order-sensitive missing

\* 基金项目: 国家自然科学基金(61472071, 61272179); 国家重点基础研究发展计划(973)(2012CB316201); 中央高校基本科研业务费(N140404013)

Foundation item: National Natural Science Foundation of China (61472071, 61272179); National Key Basic Research Program of China (973) (2012CB316201); Fundamental Research Funds for Central Universities (N140404013)

收稿时间: 2015-09-25; 修改时间: 2016-01-12; 采用时间: 2016-02-22

value imputation framework called OMSMVI for multi-source sensory data. OMSMVI takes advantages of multi-dimensions relevancy, such as temporal relevancy, spatial relevancy and attributive relevancy of sensing data adequately. The missing-sources-centered similarity graphs are constructed based on multi-dimensions relevancy. At the same time, in the process of missing data imputation, the imputed missing values are used as observations to impute subsequent missing values. Taking the whole distribution of missing sources into consideration, the framework performs order-sensitive missing value imputation, meaning that the order of imputation is ascertained before applying the specific MVI (missing value imputation) methods. Order-sensitive imputation can remit the decrease of imputed result accuracy caused by the lower similarity between missing source and its neighbors when the missing sources are dense. Finally, a new neighborhood-based missing values imputation algorithm NI, which modifies the KNN imputation algorithm, is introduced into the OMSMVI framework. NI uses the multi-dimension similarity to search the missing sources' neighbors which reflect the similarity from multiple dimensions. Such NI algorithm overcomes the shortcoming that parameter K of KNN is difficult to determine. Furthermore, NI algorithm can improve the imputation accuracy further compared to KNN. Two true sensor data sets are used to compare with the baseline MVI methods to verify the accuracy and effectiveness of OMSMVI.

**Key words:** missing values; dense missing; sensing network; sequential-sensitive imputation; multi-dimensions relevancy

由于感知设备硬件资源有限、抵制干扰性差等固有限制,感知网络在数据获取过程中经常存在数据缺失现象.目前,一般对缺失数据的处理办法分为 3 大类:(1) case deletion,即直接丢弃缺失数据元组;(2) learning without handling of missing data,即不做处理,直接将缺失数据元组传递给上层应用;(3) data imputation,对缺失数据进行填补.当数据集中包含大量缺失数据时,若直接将其丢弃会造成原始数据信息的失真,导致上层应用得到的结果与预期偏差较大.若不对缺失数据进行处理,现有的用于聚类(clustering)、统计分析(statistical analysis)、机器学习(machine learning)等上层应用中的分析工具无法处理含有缺失值的数据集.因此,当缺失数据的规模对整个数据集来说不可忽视时,对缺失数据进行填补是非常必要的.

由于感知网络主要通过部署大量的感知设备对真实环境中的目标从多个维度进行监测,因此,感知数据通常具有一定的时间相关性、空间相关性以及属性相关性,本文称这 3 方面相关性为多维度相关性.目前,已有的缺失数据填补算法本质上均利用了数据在时间或空间或属性上的相关性,如:连续多元回归填补<sup>[1-4]</sup>通过衡量数据源在空间或属性上的相似度来获取用于估计缺失值的数据元组;时间序列填补<sup>[5]</sup>利用数据的时间相关性进行填补;加权  $k$  近邻填补<sup>[6]</sup>利用数据的属性相关性进行填补.但这些已有方法均未同时考虑感知数据的多维度相关性,而只考虑某一方面的相关性通常具有一定的局限性.

例 1:图 1 是伯克利研究实验室提供的室内 54 个传感器的分布图,利用这 54 个传感器可对室内温度、湿度和光强数据进行采集.

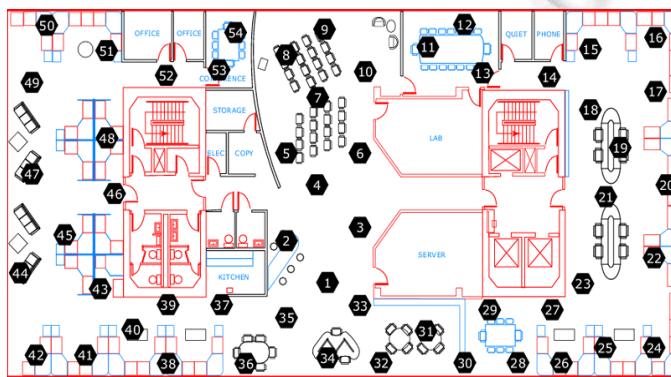


Fig.1 Distribution of 54 real sensors

图 1 54 个真实传感器的分布图

表 1 为图 1 中数据源  $s_{42} \sim s_{51}$  在  $t_1 \sim t_7$  时刻对室内温度、湿度信息进行获取得到的感知数据矩阵.表中第 1 行代表时间戳,第 1 列代表数据源,其余单元格内为相应数据源在对应时刻的温度、湿度感知数据对,加粗的数字的表示该数据在真实数据集中是缺失的.根据各个传感器的位置分布可知,传感器  $s_{46}$  的空间 3NN 为  $s_{45}, s_{47}, s_{48}$ ,

但  $t_7$  时刻  $s_{45}, s_{47}, s_{48}$  的温度数据均为缺失状态,即:若只利用空间相关性,将无法对传感器  $s_{46}$  进行填补.而根据各个感知节点的历史数据,可计算得到传感器  $s_{46}$  在时间相关性上的  $3NN$  为  $s_{47}, s_{50}, s_{51}$ .由于  $s_{50}, s_{51}$  在  $t_7$  时刻为完整数据元组,可利用时间相关性上的近邻节点对传感器  $s_{46}$  的温度值进行填补.因此,同时考虑感知数据的多维度相关性,将更加有利于缺失数据的填补,且 3 个方面的相关性具有相辅相成的作用.

**Table 1** Sensing data of sources  $s_{42} \sim s_{51}$  at  $t_1 \sim t_7$

**表 1** 数据源  $s_{42} \sim s_{51}$  在  $t_1 \sim t_7$  的感知数据

	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	$t_6$	$t_7$
$S_{42}$	(19.037,43.452)	(18.814,43.428)	(18.466,43.543)	(18.202,43.910)	(18.012,44.412)	(17.651,44.922)	(17.248,45.731)
$S_{43}$	(19.079,44.113)	(18.763,44.365)	(18.425,44.611)	(18.173,44.722)	(18.033,44.855)	(17.744,45.205)	(17.406,45.755)
$S_{44}$	(18.904,45.375)	(18.536,45.833)	(18.149,46.226)	(17.862,46.476)	(17.745,46.606)	(17.347,47.264)	(17.010,47.847)
$S_{45}$	(19.115,44.092)	(18.777,44.409)	(18.418,44.741)	(18.187,44.789)	(18.046,44.846)	(17.755,45.206)	(17.421,45.736)
$S_{46}$	(18.361,45.382)	(18.01,45.769)	(17.719,46.05)	(17.413,46.130)	(17.231,46.288)	(16.838,46.988)	(16.516,47.515)
$S_{47}$	(18.045,47.046)	(17.725,47.431)	(17.466,47.56)	(17.128,47.674)	(16.894,47.948)	(16.586,48.427)	(16.379,48.652)
$S_{48}$	(19.143,43.886)	(18.838,44.100)	(18.5951,44.16)	(18.319,44.272)	(18.110,44.463)	(17.831,44.829)	(17.593,45.095)
$S_{49}$	(18.997,44.902)	(18.74,44.936)	(18.463,45.069)	(18.211,45.213)	(18.045,45.248)	(17.785,45.637)	(17.451,46.239)
$S_{50}$	(18.441,46.305)	(18.148,46.455)	(17.859,46.619)	(17.596,46.751)	(17.434,46.790)	(17.162,47.233)	(16.959,47.538)
$S_{51}$	(18.261,46.434)	(17.874,46.794)	(17.612,46.918)	(17.376,47.032)	(17.213,47.076)	(17.47.366)	(16.853,47.542)

另一方面,与待填补数据源相似度较高的数据源是否包含缺失数据,将对填补值的准确性有很大的影响,尤其是在缺失数据较为密集的情况下.因此,本文将已填补的缺失值作为观测值用于后续填补,此时,不同的缺失数据填补顺序将得到不同甚至差异很大的填补结果.

例 2:在对表 1 中的缺失数据进行填补过程中,根据不同的填补顺序会得到不同的填补结果,如填补顺序为 (46, 48,45,49,44,47) 时,填补结果为 (17.10,17.214,16.905,16.904,17.291,17.22);当填补顺序为 (49,47,48,46,44,45) 时,填补结果为 (17.151,17.248,16.906,16.964,17.327,17.248).由此可知,不同填补顺序得到的填补结果不同.此外,根据表 1 给出的各个缺失值的真值,可计算得到第 1 个填补顺序的平均填补误差为 0.290,第 2 个填补顺序的平均填补误差为 0.293.由此说明,不同填补结果对应不同的填补误差.如何确定一个最优的填补顺序,从而得到填补误差最小的填补结果,是本文重点要解决的问题.

基于以上描述,本文提出了面向多源感知数据的、基于多维度相关性且顺序敏感的缺失值填补框架 OMSMVI.该框架首先利用感知数据的多维度相关性,提出对缺失数据源与其他数据源间相似度度量的新方法;其次,OMSMVI 构建能够充分刻画缺失数据源之间以及缺失数据源与完整数据源间相关关系的相似图,进而在相似图基础上提出缺失数据填补顺序的决策算法;最后对 KNN 填补算法进行改进,提出一种新的基于近邻的缺失值填补方法 NI.本文的主要贡献点主要有:

- 面向感知数据的缺失值填补问题,提出一种新的顺序敏感的缺失值填补框架 OMSMVI;
- 充分利用感知数据的多维度相关性对数据源间的相似度进行衡量,并构建以缺失数据源为中心的相似图;
- 针对缺失数据较为密集的情况,考虑不同填补顺序对填补精度的影响,提出填补顺序决策问题,并通过精确和近似算法对其进行求解;
- 对现有的 KNN 填补算法进行改进,提出基于近邻的缺失值填补算法,该算法解决了 KNN 填补过程中  $K$  值难以确定的问题;并通过利用缺失数据源周围的所有近邻节点对缺失值进行填补,提高了在缺失值较为密集情况下填补值的准确性;
- 利用两个真实数据集,通过大量实验验证了本文算法的准确性及有效性.

本文第 1 节对已有的缺失数据填补算法进行总结,并指出已有工作与本文的不同.第 2 节给出问题描述及相关定义.第 3 节详细介绍缺失数据源的近邻筛选方法.第 4 节提出缺失数据填补顺序决策问题,给出精确及近似求解方法,并提出一种新的缺失值填补算法 NI.第 5 节给出实验结果及数据分析.最后,在第 6 节进行总结.

## 1 相关工作

目前已有的填补算法很多,常用的算法有 hot-deck 填补、回归填补、KNN 填补、多重填补。Hot-deck 填补主要包含两个步骤:(1) 对不包含缺失值且与待填补值相关的完整数据进行分类;(2) 从每个分类中按照完整数据与缺失数据的相关性大小选择合适的完整数据作为 donor 用于填补<sup>[7]</sup>。文献[8]在文献[7]的基础上,在 donors 选取计算中加入权重,并提出每个 donor 用于缺失值填补的次数与权重成正比。回归填补是一种条件性的均值填补,该方法基于完整数据集建立缺失值与已知值间的回归模型,进而依据历史数据学习得到的参数来估计缺失变量值。当变量不是线性相关或高度相关时,会导致填补值与真实值间的偏差较大。文献[3,4]利用一系列线性和非线性回归模型对缺失值进行填补,文献[1,2]利用核函数建立回归模型。回归填补的优点是既适用于分类型数据又适用于连续型数据,缺点是模型的参数确定比较复杂,填补准确性不容易保证。KNN 填补的主要思想是,利用缺失值的  $K$  个近邻的加权均值替代缺失值。文献[6]针对基因数据,利用欧式距离找出数据矩阵中与缺失值最相似的  $K$  个近邻进而进行缺失值填补。文献[9]基于灰色关联度对缺失数据的  $K$  近邻进行查找,进而对缺失值进行填补。文献[10]提出基于 KNN 的部分填补,该方法利用机器学习和数据挖掘异常点检测技术来探测缺失值是否可填补:若缺失值不可填补,则放弃填补;若缺失值可填补,则利用缺失值的左右近邻进行加权 KNN 填补。由此可知,该填补方法并不是对每个缺失值均进行填补。为了进一步提高填补的精度,文献[11]提出了基于 EM 的非参数缺失数据填补方法,类似于 EM 算法,不同的是,利用非参数模型如 KNN 或核函数回归来代替 EM 算法中的有参模型。多重填补<sup>[12-14]</sup>利用  $m(m>1)$  个能够反映数据本身分布概率的值来代替缺失值,以反映缺失数据的不确定性。多重填补通过模拟缺失数据的分布,较好地保持了变量之间的关系,但其处理过程较为复杂。

以上算法均针对静态数据进行填补,文献[15,16]研究了感知流上的缺失值估计问题。Gruenwald 等人<sup>[15]</sup>采用数据挖掘技术,提出了 WARM 算法,当某一数据源节点  $a$  产生的流数据出现缺失时,该算法首先找到与  $a$  相关联的另一数据源节点  $b$ ,然后用  $b$  的相应数据作为  $a$  的填补值。Jiang 等人<sup>[16]</sup>对 WARM 算法进行了改进,提出了 CARM 算法,该算法通过对流数据进行关联规则计算,找到多个数据源节点的频繁模式,并用该频繁模式来估计缺失值。然而这两个算法具有很大的局限性,无法被广泛应用:首先,这两个算法只能处理离散型数据,无法处理连续型数据;另一方面,这两个算法对缺失值的计算能力和估计准确性依赖于关联规则中用户指定的支持度阈值和置信度阈值,而实际应用中,用户很难知道数据之间的关联程度及数据变化规律,因此很难给出恰当的阈值。此外,如果缺失值所对应的数据元组不出现在频繁模式中,则这两个算法无法对缺失值进行估计计算。文献[17]提出了针对缺失数据流进行并行填补的框架 Pythia,该框架主要解决的问题是:如何将包含缺失值的大数据划分到多台机器上进行并行处理,从而提高填补效率。框架用到的填补方法为已有的加权 KNN 填补及连续多元回归填补,因此侧重点和本文不同。

另一方面,由于数据类型限制或处理速度的制约,目前针对感知数据,同时考虑其在时间、空间、属性这 3 方面相关性的缺失值填补方法鲜少。文献[18]基于稀疏真实数据,利用机器学习方法对细粒度的空气质量数据进行填补。该工作从大量交通流量、气象等数据集中提取出与空气质量相关的特征属性,并将这些特征属性分为空间相关的和时间相关的两类,进而分别对其利用半监督机器学习方法并结合已有的历史数据进行缺失值的填补。该方法的局限性在于:只能处理分类型数据,针对连续型数据,只能先对其进行离散化处理才能在机器学习过程中对其进行打标签操作,因此不适用于连续型数据。

最后,以上相关工作均未考虑在缺失值较为密集的情况下,不同的填补顺序对填补结果的影响。目前,针对顺序敏感的缺失数据填补的方法还比较少, Kim 等人<sup>[19]</sup>提出了顺序 KNN 填补算法,该算法针对 gene 数据,根据每个基因中包含缺失值的多少对其进行排序,先填补包含缺失值最少的数据,并将填补后的 gene 作为完整数据用于后续缺失值的填补。算法只是通过简单的规则定义缺失值的填补顺序,并未对不同填补顺序对填补结果的影响进行量化。文献[20]在确定填补顺序的策略方面与文献[19]完全相同,只是在判定不同 gene 间相似度的方法上不同。

## 2 问题描述

设感知网络中有  $N$  个数据源  $S = \{S_1, S_2, \dots, S_N\}$ ,  $t$  时刻  $N$  个数据源的感知数据集合为  $\mathcal{D}^t = \{D_1^t, D_2^t, \dots, D_N^t\}$ , 其中每条感知数据为包含  $m$  维属性的多模态数据, 如数据源  $S_i$  在  $t$  时刻的感知数据为  $D_i^t = (d_{i1}^t, d_{i2}^t, \dots, d_{im}^t)$ . 下面是文中要用到的一些定义.

**定义 1(缺失数据元组).** 若多模态感知数据  $D_i^t$  在某一维上的属性值缺失, 则称  $D_i^t$  为缺失数据元组.

**定义 2(完整数据元组).** 若多模态感知数据  $D_i^t$  在每一维属性上观测值均存在, 则称  $D_i^t$  为完整数据元组.

**定义 3(相对完整数据元组).** 对于缺失数据  $D_i^t$  的待填补维  $d_{ij}^t$ , 若多模态感知数据  $D_k^t$  在待填补维上的观测数据  $d_{kj}^t$  存在, 则称  $D_k^t$  为  $D_i^t$  的相对完整数据元组.

**定义 4(时间邻域).**  $t$  时刻, 对于待填补缺失数据  $d_{ij}^t$ , 其前  $s$  个时刻的感知数据元组集合称为  $d_{ij}^t$  的时间邻域, 表示为  $\mathcal{D}_{ij}^{TN} = \{D_{ij}^{t-s}, D_{ij}^{t-(s-1)}, \dots, D_{ij}^{t-1}\}$ .

**定义 5(属性邻域).**  $t$  时刻, 对于待填补缺失数据  $d_{ij}^t$ , 其属性邻域为数据元组  $D_i^t$  中与待填补维  $j$  具有相关性的属性组成的新的元组, 表示为  $\mathcal{D}_{ij}^{AN} = \{d_{ik}^t \mid k \in [1, m], k \neq j\}$ .

**定义 6(空间近邻).**  $t$  时刻, 数据源  $S_i$  的空间近邻为与其欧式距离小于等于给定阈值  $\delta_d$  的数据源集合, 表示为  $SN_i = \{S_j \mid j \in [1, N], j \neq i, dis(S_i, S_j) \leq \delta_d\}$ .

为了简便起见, 不失一般性, 本文利用数据元组中只包含一维数据缺失的情况进行说明. 对于包含多维缺失数据的数据元组, 可将其看作多条只包含一维缺失数据的数据元组进行处理. 设  $t$  时刻到达的  $N$  个数据源的感知数据集中包含  $n$  条缺失数据元组, 对应的缺失数据源集合为  $MS = \{S_1, S_2, \dots, S_n\}$ . 本文最终的目的是要对  $n$  个缺失数据源的缺失数据进行填补, 利用的基本思想是: 对于缺失数据源  $S_i$ , 若用于填补的数据源与缺失数据源间的相似度越高, 则填补值的准确性越高; 同时, 两个缺失数据源间的相似度越高, 则在数据填补过程中, 他们之间的相互影响越大. 由于每个缺失数据源的近邻分布不同, 在缺失数据分布较为密集的情况下, 不同的填补顺序会得到不同的填补结果, 因为已填补的缺失值将作为观测值用于后续填补过程. 因此, 在对缺失数据进行填补前, 需确定最优的填补顺序. 本文提出的顺序敏感的缺失值填补框架 OMSMVI 的基本流程如图 2 所示.

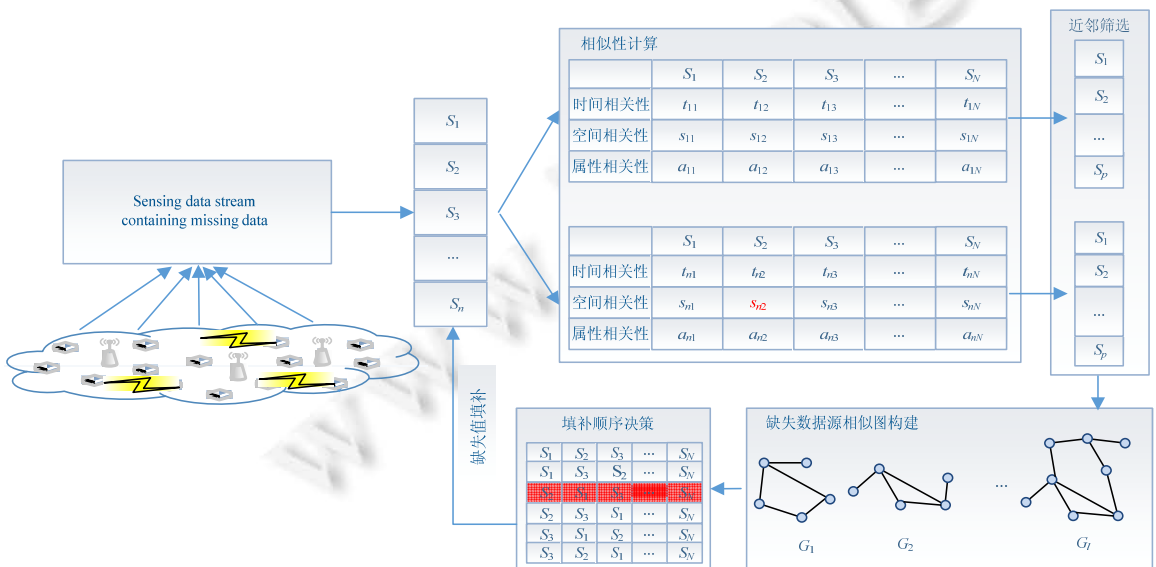


Fig.2 Sequential imputation framework of OMSMVI

图 2 顺序填补框架 OMSMVI

对于  $t$  时刻感知数据中包含的  $n$  条缺失数据,首先计算每个缺失数据源与剩余数据源在时间、空间、属性上的相关性,进而确定数据源间的多维度相关性;其次,依据数据源间的多维度相似性确定每个缺失数据源的近邻节点,进而基于近邻节点构建以缺失数据源为中心的相似图,由于相似图中有边连接的两点间的相似度需大于给定阈值,因此构建出的相似图可能是一个包含所有缺失数据源的大图,也可能是多个互不连通的小图;最后,基于相似图对缺失数据的填补顺序进行决策,选择最优的填补顺序,进而利用适用的缺失数据填补方法对缺失数据依次进行填补。

### 3 缺失数据源近邻筛选

本节将详细介绍缺失数据源的近邻节点筛选算法,其主要思想是:综合缺失数据源与剩余数据源在时间、空间、属性这 3 方面的相似度,筛选出每个缺失数据源的近邻节点,近邻节点与缺失数据源间的相似度需在某一维度上大于相应阈值。为了能最大限度地找出所有近邻点,时间、空间、属性这 3 方面的相关性相互互补,即,近邻节点只需满足某一方面的相似度即可。

#### 3.1 多维度相关性模型

多维度相关性是指感知数据特有的时间相关性、空间相关性以及属性相关性。时间相关性是指感知数据随时间的变化具有一定的趋势性,若两个数据源在历史时间内的数据比较相似,则下一时刻二者之间的观测值相差不大,即: $t$  时刻对第  $j$  维缺失数据进行填补过程中,若数据源  $S_i$  和  $S_k$  的时间邻域  $\mathcal{D}_{ij}^{TN}, \mathcal{D}_{kj}^{TN}$  相似度较高,则  $d_{ij}^t$  与  $d_{kj}^t$  也较为相似。本文利用数据源在时间邻域上的相似度来衡量当前时刻两数据源间的时间相似性。感知数据的空间相关性是指物理位置上相近的感知节点采集到的数据往往比较相似或存在某种函数关系。直观上,若两个数据源相距很远,它们将不具有空间相关性。因此对于缺失数据源  $S_i$ ,只需计算其空间近邻内的数据源与  $S_i$  的相似性即可。此外,由于数据源的位置通常是相对固定或是成组移动的,因此可周期性地对各个数据源间的空间相似性进行更新,而不必每一时刻均对其进行计算。属性相关性是指不同维度间的感知数据具有一定的相关性,如海水的温度和盐度、空气的能见度和风速均存在一定的正相关关系。在  $t$  时刻对第  $j$  维缺失数据进行填补过程中,若  $d_{ij}^t$  与  $d_{kj}^t$  的属性邻域具有较高的相似度,则  $d_{ij}^t$  与  $d_{kj}^t$  相差不大。基于此,本文通过待填补维度的属性邻域的相似性来衡量不同数据源在待填补维上的属性相似性。

综上,基于对不同数据源在时间、空间、属性上相似性度量的介绍,本文采用平均欧式距离来度量数据源间多维度的相似度。假设当前时刻第  $j$  维属性存在数据缺失,则数据源  $S_i$  和  $S_k$  间的多维度相似度为

$$Sim_{ik} = \frac{1}{1 + \min\{d_{ik}^T, d_{ik}^S, d_{ik}^A\}} \tag{1}$$

其中,  $d_{ik}^T, d_{ik}^S, d_{ik}^A$  分别表示数据源  $S_i$  和  $S_k$  在时间、空间、属性上的距离,具体计算公式如下所示:

$$d_{ik}^T = \frac{h_T}{s} \sqrt{\sum_{d_{ij}^t \in \mathcal{D}_{ij}^{TN}, d_{kj}^t \in \mathcal{D}_{kj}^{TN}} (d_{ij}^t - d_{kj}^t)^2} \tag{2}$$

$$d_{ik}^S = \frac{h_S}{|loc|} \sqrt{\sum_{l_i \in loc_i, l_j \in loc_j} (l_i - l_k)^2} \tag{3}$$

$$d_{ik}^A = \frac{h_A}{|\mathcal{D}_{ij}^{AN}|} \sqrt{\sum_{d_{ij}^t \in \mathcal{D}_{ij}^{AN}, d_{kj}^t \in \mathcal{D}_{kj}^{AN}} (d_{ij}^t - d_{kj}^t)^2} \tag{4}$$

其中,  $loc$  表示数据源的位置信息;  $|loc|$  表示位置的维数,如感知设备位置信息用二维空间点表示,则  $|loc|=2$ ;  $|\mathcal{D}_{ij}^{AN}|$  为属性邻域的维数,  $h_T, h_S, h_A$  为每一维度上的归一化因子。根据公式(1)可知,两个数据源间的多模态相似度为它们在时间、空间、属性上相似度的最大值。这样的定义能够保证在缺失数据源近邻节点筛选过程中更多的节点被保留,并保证相似度值最大。也由此说明感知数据在时间、空间、属性这 3 方面的相似度相互互补,从而缓解在缺失数据分布较密集情况下近邻节点均为数据缺失导致无法给出填补值的情况。此外,对于相似性度量方法

的选取并不局限于欧式距离,可根据具体应用选择合适的相似度度量方法.

### 3.2 缺失数据源近邻节点筛选

根据第 3.1 节介绍的多维度相关性模型,可求得每个缺失数据源与剩余数据源间的多维度相似度.由于缺失数据源的近邻节点将用于填补顺序的确定以及缺失值填补过程中,而相似度较低的节点对填补顺序及填补值的确定影响较小,为了降低相似图构建的计算量,首先对每缺失数据源的近邻节点进行筛选,筛选规则很简单,即,相似度小于给定多维度相似度阈值 $\delta_s$ 的节点将被删除.

例 3:仍以表 1 中给定数据为例,根据第 3.1 节对多维度相似度计算的定义,可得到每个缺失数据源与剩余数据源的多维度相似性见表 2,假设 $\delta_s=0.8$ ,则每个缺失数据源的近邻节点列表如图 3 所示.

**Table 2** Multimodal similarity between missing sources and all sources of Table 1

表 2 表 1 中各缺失数据源与剩余数据源间的多维度相似度

	$S_{42}$	$S_{43}$	$S_{44}$	$S_{45}$	$S_{46}$	$S_{47}$	$S_{48}$	$S_{49}$	$S_{50}$	$S_{51}$
$S_{44}$	0.897	0.894	-	0.891	0.92	0.835	0.862	0.889	0.881	0.818
$S_{45}$	0.995	0.992	0.891	-	0.755	0.694	0.959	0.977	0.800	0
$S_{46}$	0.761	0.759	0.92	0.755	-	0.894	0.735	0.757	0.977	0.973
$S_{47}$	0.794	0.697	0.836	0.694	0.894	-	0.698	0.696	0.838	0.901
$S_{48}$	0.953	0.956	0.862	0.959	0.735	0.698	-	0.958	0.698	0.727
$S_{49}$	0.973	0.982	0.889	0.977	0.757	0.696	0.958	-	0.803	0.749

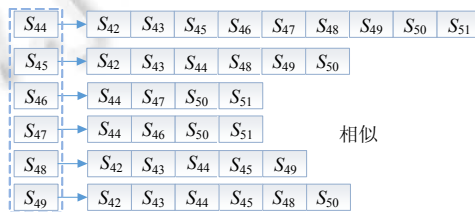


Fig.3 Neighboring sources list of missing sources

图 3 缺失数据源的近邻节点列表

## 4 缺失数据填补顺序决策及缺失值填补

在缺失数据分布较为密集的情况下,缺失数据源的近邻节点也可能数据缺失,此时,利用已有的缺失数据填补方法会导致填补值的准确性降低,甚至无法对缺失值进行填补.基于此,本文将已填补的缺失值作为观测值用于后续填补过程,但填补值具有一定的不确定性,当前填补值的准确性将直接影响后续与其相关的填补值的精度,由此便引发填补顺序的问题.如引言中例 2 的介绍,不同的填补顺序会得到不同甚至差异很大的填补结果,因此,如何确定一个最优的填补顺序以提高填补值的准确性,成为一个具有挑战性并亟待解决的问题.本文首先将缺失数据源集合构建为一个相似图结构,进而在相似图基础上,将填补顺序决策问题转化为一个最优化问题并对其进行求解.

### 4.1 缺失数据源相似图构建

根据第 3 节对缺失数据源近邻节点筛选策略的介绍,可得到  $t$  时刻  $n$  个缺失数据源的近邻节点列表以及缺失数据源与各个近邻节点间的相似度,本节将根据这些数据信息构建以缺失数据源为中心的相似图.由于相似图的构建主要用于缺失数据源填补顺序的确定,因此,图的顶点只包含缺失数据源.顶点权重为该缺失数据源与其所有相对完整近邻节点间相似度融合后的结果.边表示两个缺失数据源互为近邻节点,边权重为二者的相似度.由此构建出能够直观反映缺失数据源间依赖关系的无向加权相似图,并且通过边权重反映两个缺失数据源间相似度的大小,通过顶点权重反映每个缺失数据源的所有相对完整近邻节点的分布情况,计算公式见公式(5).

$$w_{s_i} = 1 - \prod_{s_k \in N(s_i)} (1 - Sim_{ik}) \tag{5}$$

其中,  $N(s_i)$  表示缺失顶点  $s_i$  的相对完整近邻节点集合,  $Sim_{ik}$  为数据源  $s_i$  与  $s_k$  间的多维度相似性. 根据公式(5)可知: 缺失数据源  $s_i$  的相对完整近邻节点个数越多, 且每个近邻节点与  $s_i$  的相似度越大, 则  $s_i$  顶点的权重值越大.

例 4: 以例 3 中缺失数据源近邻点的计算结果为例进行相似图构建, 得到的缺失数据源相似图如图 4(a) 所示.

需要注意的是:  $t$  时刻, 由  $n$  个缺失数据源构建出的相似图可能是一个包含所有节点的大图, 也可能是多个互不连通的子图. 例如: 当例 3 中的多维度相似性阈值  $\delta_s=0.9$  时, 缺失数据源相似图则由 3 个子图构成(如图 4(b) 所示). 在对缺失数据源进行填补顺序确定过程中, 每个子图之间是相互独立的.

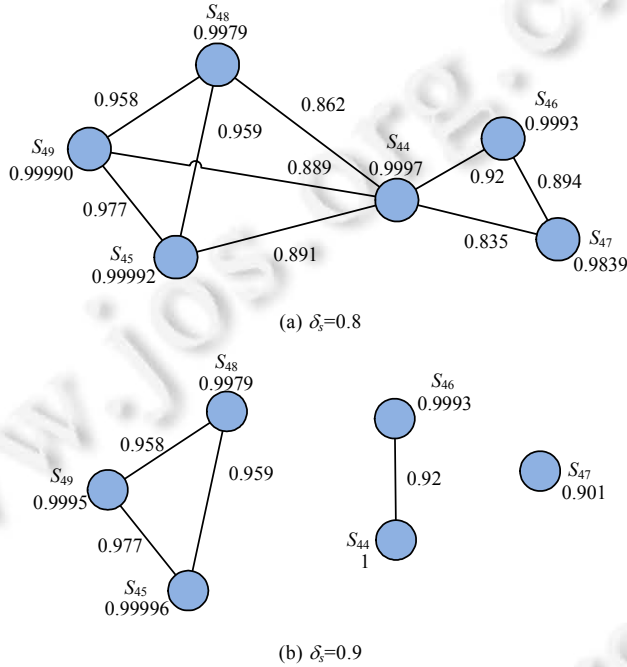


Fig.4 Similarity graph of missing sources

图 4 缺失数据源相似图

### 4.2 填补顺序决策

基于第 4.1 节构建的缺失数据源相似图, 本节主要要解决的问题是对缺失数据源填补顺序的决策. 根据第 4.1 节, 相似图的顶点权重反映的是可用于填补该缺失数据源的相对完整近邻节点的个数以及近邻节点与之的相似度. 权重值越大, 表明其周围可用于填补的相对完整数据越多且相似度越高, 进而反映了其缺失近邻节点在填补过程中的贡献量相对越小. 而边权重反映的是两个缺失数据源间的相似度, 该相似度越高, 表明数据填补过程中两个数据源间的相互影响越大.

当给定一个填补顺序时, 每对互为近邻节点(即由一条边相连)的两个缺失数据源的填补优先级即可被确定, 进而相应的无向加权相似图即可转化为一个有向加权相似图.

例 5: 针对例 4 中的缺失数据源相似图 4(a), 假设填补顺序为(44,46,47,49,48,45), 则图 4(a)可转化为一个有向加权图, 如图 5 所示.

对于一个包含  $n$  个缺失数据源的可能的填补顺序  $seq$ , 本文将对应的有向加权相似图看作一个贝叶斯网络, 顶点权重为贝叶斯网络中每个顶点的先验概率, 边权重为两个顶点间的条件概率, 该贝叶斯网络的联合概率为相应填补顺序的置信度, 计算公式见公式(6).

$$b(seq) = \prod_{i=1}^n p(s_i | s_{N(i)}) \tag{6}$$



其中,  $s_{N(i)}$  为可用于缺失数据源  $s_i$  填补的缺失近邻节点集合. 当一个缺失数据源被填补后, 可将其视为观测值并用于后续缺失值填补过程中. 因此在填补顺序确定后, 后填补的缺失数据源的近邻节点是动态变化的, 顶点的权重也是动态更新的,  $p(s_i | s_{N(i)})$  为顶点  $s_i$  在可用于填补的近邻节点列表中加入  $s_{N(i)}$  后更新的顶点权重值. 若  $s_{N(i)}$  为空, 则  $p(s_i | s_{N(i)})$  可表示为  $p(s_i)$ , 即, 缺失数据源  $s_i$  的初始顶点权重  $p(s_i) = w_{s_i}$ ; 若  $s_{N(i)}$  不为空, 结合公式(5), 可得  $p(s_i | s_{N(i)})$  见公式(7).

$$p(s_i | s_{N(i)}) = 1 - (1 - w_{s_i}) \prod_{s_j \in N(i)} (1 - Sim_{ij}) \tag{7}$$

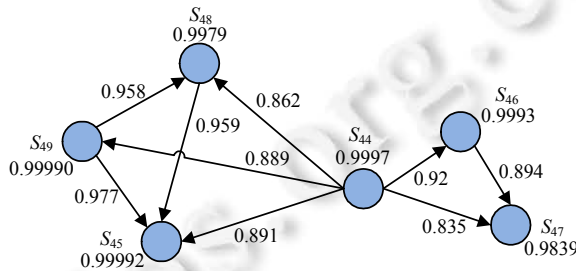


Fig.5 Similarity digraph of missing sources ( $\delta_s=0.8$ )  
图 5 有向缺失数据源相似图( $\delta_s=0.8$ )

**引理 1.** 对于缺失数据源  $s_i$ ,  $p(s_i | s_{N(i)})$  越大, 填补误差越小.

证明: 目前已有的基于统计学的缺失值填补算法均遵循一个共同的填补思想: 通过与缺失数据源较为相似的完整数据源对其进行填补, 并认为完整数据源的个数越多且与缺失源间的相似度高, 则填补精度越高. 本引理将基于此基本填补思想给出证明.

根据  $p(s_i | s_{N(i)})$  的定义可知:  $s_{N(i)}$  为空时,  $p(s_i | s_{N(i)}) = p(s_i) = w_{s_i}$ . 结合公式(5)可知:  $w_{s_i}$  反映的是缺失数据源的完整近邻节点分布, 完整近邻节点与缺失节点间的相似度高且完整近邻节点个数越多, 公式(5)的后半部分越小, 即  $p(s_i | s_{N(i)})$  越大, 则根据缺失值的基本填补思想, 其填补误差越小.  $s_{N(i)}$  不为空时, 根据公式(7)可知: 可用于缺失值填补的填补近邻节点个数越多且二者间的相似度高,  $\prod_{s_j \in N(i)} (1 - Sim_{ij})$  越小, 相应的  $p(s_i | s_{N(i)})$  越大, 则填补误差越小.

综上, 引理 1 得证. □

**定理 1.** 给定一个填补顺序, 其置信度越高, 填补误差越小.

证明: 对于一个给定的填补顺序, 根据公式(6)的定义可知:  $p(s_i | s_{N(i)})$  越大, 该填补顺序对应的置信度越高. 再结合引理 1, 不难得出定理 1 的结论. □

例 5: 基于构建好的相似图, 可通过枚举计算每个可能填补顺序的置信度, 表 3 显示的是与例 4 中图 4(a) 相似图对应的部分填补顺序及置信度以及填补误差.

**Table 3** Part of possible imputation sequences and their confidence and average error  
**表 3** 部分填补顺序及其置信度、平均填补误差

填补顺序	置信度	平均填补误差
(44,46,47,49,48,45)	0.587 5	0.330 6
(49,46,48,47,45,44)	0.634 9	0.289 8
(45,49,46,48,44,47)	0.632 0	0.307 9
(45,48,47,44,46,49)	0.600 1	0.303 8
(48,45,44,49,47,46)	0.598 4	0.318 3
...	...	...

从表中可见: 不同填补顺序的置信度是不同的, 填补误差也不同, 置信度越高的填补顺序, 填补误差越小.

基于以上介绍, 最优填补顺序决策问题即转化为置信度最高的贝叶斯网络选取问题. 设  $n$  个缺失数据源所

有可能的填补顺序集合为  $SEQ = \{seq_1, seq_2, seq_3, \dots\}$ , 填补顺序  $seq_i$  的置信度为  $b(seq_i)$ , 则填补顺序决策问题即转化为求  $SEQ$  中置信度最大的序列:

$$seq^* = \operatorname{argmax}(b(seq_i)), seq_i \in SEQ.$$

对于上述问题, 可通过枚举所有可能的填补顺序, 选取置信度最大的序列作为最优填补顺序得到最优解, 见算法 1.

**算法 1.** 填补顺序决策精确算法.

输入: 缺失数据源相似图集合  $\mathcal{G} = \{G_1, G_2, \dots\}$ ;

输出: 缺失数据源最优填补顺序  $seq^*$ .

1. **foreach** similarity graph  $G_i \in \mathcal{G}$
2.   all possible imputation sequences  $SEQ_i$  is
3.    $SEQ_i = \text{Enumerate}(G_i) = \{seq_1, seq_2, \dots\}$
4.   **foreach** imputation sequence  $seq_j \in SEQ_i$
5.     calculate its confidence  $b(seq_j)$  according to Eq.(6)
6.   end
7.   select the best imputation sequence  $seq_i^*$  whose confidence is the maximum, and then  
 $seq^* = seq_1^* \cup seq_2^* \cup \dots$
8. end
9. **return**  $seq^*$

根据算法 1 可知, 最优填补顺序精确解法的时间复杂度为指数级  $O(n!)$ ,  $n$  为一个相似图中缺失数据源的个数. 当  $n$  较小时, 可通过枚举法得到精确的最优填补顺序; 但当数据量较大且缺失数据分布较为密集(即  $n$  较大)时, 算法的复杂度将呈指数级增长, 精确求解问题也随之成为 NP 难问题. 为了降低算法的时间复杂度, 本文利用启发式算法对其进行近似求解. 算法的基本思想是: 对于两个具有依赖关系的缺失数据源来说, 先填补顶点标签值较高的缺失数据源得到的最终填补结果相对较为准确. 这是因为顶点标签值越高, 表明该缺失数据源的近邻完整数据源与之的相似度越高, 也就意味着填补值相对来说更加准确. 基于此, 本文采用启发式算法对最优填补顺序进行决策, 主要包含 3 个步骤.

- 1) 对每个相似图中的顶点标签值进行排序, 选择顶点标签值最高的节点作为优先填补对象;
- 2) 对该缺失数据源有边相连的顶点标签值进行更新, 因为该缺失数据源填补完成后即作为完整数据用于后续填补过程, 也就是说, 该已完成填补的缺失数据源对于与之相连的缺失数据源来说是完整数据元组;
- 3) 重复步骤 1) 进行下一次最优填补对象选取.

具体算法见算法 2. 根据算法 2 可知: 在基于启发式算法的填补顺序决策过程中, 每次最优填补对象选取的时间复杂度为  $O(n)$ , 重复  $n$  次的时间复杂度为  $O(n \log n)$ , 因为每完成一次最优填补对象的选取, 下一次选取过程中的可选择对象范围即减 1.

**算法 2.** 填补顺序决策启发式算法.

输入: 缺失数据源相似图集合  $\mathcal{G} = \{G_1, G_2, \dots\}$ ;

输出: 缺失数据源最优填补顺序  $seq^*$ .

1. **foreach** similarity graph  $G_i \in \mathcal{G}$
2.   the vertexes contained in  $G_i$  is  $V = \{v_1, v_2, \dots\}$
3.   **while** ( $V \neq \text{null}$ )
4.     Current loop max vertex confidence  $c_v^* = 0$
5.     **foreach** vertex  $v_j \in V$

6. If ( $c_{v_j} > c_{v^*}$ ), then
7.      $c_{v^*} = c_{v_j}, v^* = v_j$
8.      $seq_i^* \leftarrow \{v^*\}$
9.     Delete  $v^*$  from  $V$
10. end
11.  $seq^* \leftarrow \{seq_i^*\}$
12. end
13. return  $seq^*$

此外,根据本节对填补顺序决策的详细介绍可知:当缺失数据源构建的相似图包含多个互不连通的子图时,各个子图之间是相互独立的,没有填补顺序上的依赖关系,即:只需对每个子图确定最优填补顺序,进而将所有填补顺序进行随机组合即可.因此,最优填补顺序是不唯一的.

### 4.3 缺失数据填补

当缺失数据源填补顺序确定后,即可对缺失数据依次进行填补.由于 OMSMVIF 填补框架的核心思想是利用感知数据的多维度相关性对缺失值近邻节点进行查找,进而对缺失数据进行填补,因此在本质上,利用了数据在时间、空间、属性一方面或几方面相似性的填补方法均适用于本框架,如 KNN 填补、回归填补以及基于这两种填补的改进方法.在基于 KNN 的填补方法中,需为每个缺失数据源选取  $K$  个近邻节点用于缺失值的填补,但由于每个缺失数据源的近邻节点分布不均匀, $K$  值的确定比较困难: $K$  值过小,会导致由于填补结果对噪音敏感引起的填补精度下降; $K$  值过大,会导致近邻节点集合包含大量与待填补点相似度很低的节点,造成填补精度的下降及计算量的增加.基于回归的填补方法需通过对历史数据进行学习构建回归模型,填补结果对学习得到的参数依赖较大,且复杂度更高.基于此,本文对 KNN 填补进行改进,提出一种新的近邻填补方法 NI (neighbor-based imputation).该方法不限定  $K$  值大小,而是基于第 4.1 节构建的缺失数据源相似图,利用与待填补数据源相似度高于给定阈值的所有近邻节点进行填补.其优点是针对每个待填补数据,考虑其近邻节点的分布情况,找出适用于填补的近邻节点,而非针对所有待填补数据源找出固定的  $K$  个近邻节点.对于待填补数据源  $S_i$ ,设其近邻节点集合为  $NS_i = \{S_1, S_2, \dots, S_{|NS_i|}\}$ ,则  $S_i$  的填补值  $\hat{d}_i$  见公式(8).

$$\hat{d}_i = \frac{\sum_{S_j \in NS_i} Sim_{ij} \cdot d_j}{\sum_{S_j \in NS_i} Sim_{ij}} \quad (8)$$

此外,与已有 KNN 填补或回归填补算法不同的是,NI 填补算法中对缺失数据源近邻的查找是依据感知数据的多维度相似性,而非某一维相似性,因此,用于缺失数据源填补的近邻节点更加全面,也能够进一步提高填补值的准确性.

## 5 实验

本节将通过大量实验对算法的准确性和高效性进行了验证.采用两个真实数据集:SENSOR 和 WEATHER, SENSOR(<http://db.lcs.mit.edu/labdata/labdata.html>)数据集是由部署在 Intel Berkeley 实验室的 54 个 Mica2 传感器节点在 36 天内对室内温度、湿度、光强和节点电压每隔 30s 进行一次采集得到的连续型数据.WEATHER 数据集是由 AccuWeather(<http://www.accuweather.com/zh/us/united-states-weather>)网站提供的美国 30 个城市在一周内的天气数据,包含的监测属性有温度、湿度、气压、风速等.针对以上两个数据集,本文选取温度和湿度两个属性,并且按小时对数据进行聚合后用于本文的实验.此外,本文在 WEATHER 数据基础上生成了一组模拟数据集用于后文第 5.4 节对算法执行时间的验证.对于缺失值的选取,本文在温度属性上随机删除部分观测值,进而将其看作缺失值用于填补.此外,缺失率的计算以元组为单位,即:对于一条数据元组,若其在温度属性上的

观测值为缺失的,则称该元组为缺失数据元组.数据缺失率即为缺失数据元组条数占总元组数的比例.本文利用平均相对误差来度量填补值的准确性,假设有  $n$  个缺失值,第  $i$  个缺失值的填补值为  $\hat{d}_i$ ,真值为  $d_i$ ,则平均相对误差  $avg\_relerr$  为

$$avg\_relerr = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{d}_i - d_i}{d_i} \right|$$

本文采用 KNN 与多元线性回归(linear regression)<sup>[21]</sup>作为基本的缺失值填补算法,与本文提出的近邻填补(NI)算法进行对比.依据近似度计算中考虑到的空间或时间或属性等 3 方面的相关性,分别利用 MKNN,SKNN,TKNN,AKNN 表示考虑多维度相关性、只考虑空间相关性、只考虑时间相关性、只考虑属性相关性的 KNN 填补算法.同理,利用 MLR,SLR,TLR,ALR 和 MNI,SNI,TNI,ANI 来分别表示考虑多维度相关性和只考虑空间、时间、属性上某一维相关性的回归填补和近邻填补算法.另一方面,基于是否考虑填补顺序,可将填补算法分为顺序(sequential)填补和随机(random)填补两大类,在具体填补算法前加 S 表示顺序填补,加 R 表示随机填补,如 SMKNN 表示基于多模态相关性的顺序 KNN 填补算法,RMKNN 表示基于多模态相关性的随机 KNN 填补算法.

本文在 Inter(R) Core(TM) i7-2600 CPU @3.4GHz,8.00GB 内存和 64 位 Windows 7 操作系统运行环境下,利用 Java 语言进行实验验证.实验过程中,设定时间邻域长度  $s=10$ ,空间近邻距离阈值  $\delta_t=50$ ,近邻节点筛选过程中多模态相似度阈值  $\delta_s=0.9$ .基于 KNN 填补算法中, $K$  值取 5.这些参数的设定不是固定的,可根据具体的数据集和应用背景由用户给定.

### 5.1 单维度和多维度相关性对填补准确性的影响

根据第 3 节的介绍可知:本文利用感知数据的多维度相关性对缺失数据源近邻节点进行筛选,进而利用近邻节点对缺失值进行填补.图 6 和图 7 给出了利用多维度相关性和单维度相关性进行缺失值填补时的准确性对比.图 6(a)和图 7(a)是基于 KNN,NI,LR 的随机填补算法,图 6(b)和图 7(b)是基于 KNN,NI,LR 的顺序填补算法.

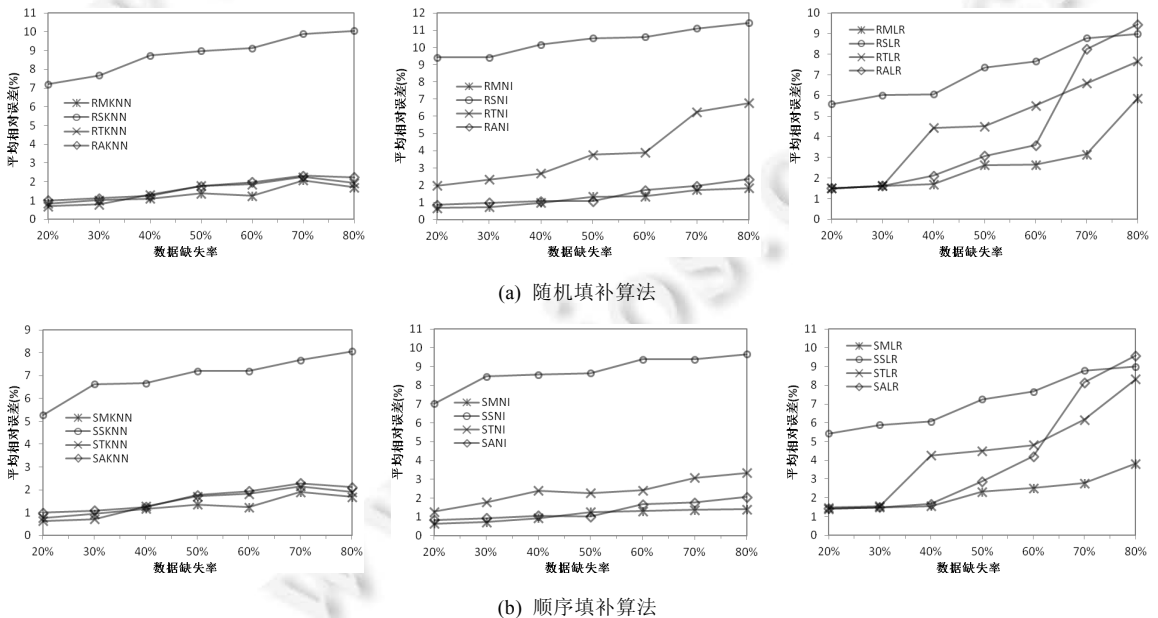


Fig.6 Influence of mulitmodal vs. single-modal on imputation accuracy(SENSOR)

图 6 多模态 vs.单模态相关性对准确性影响(SENSOR)

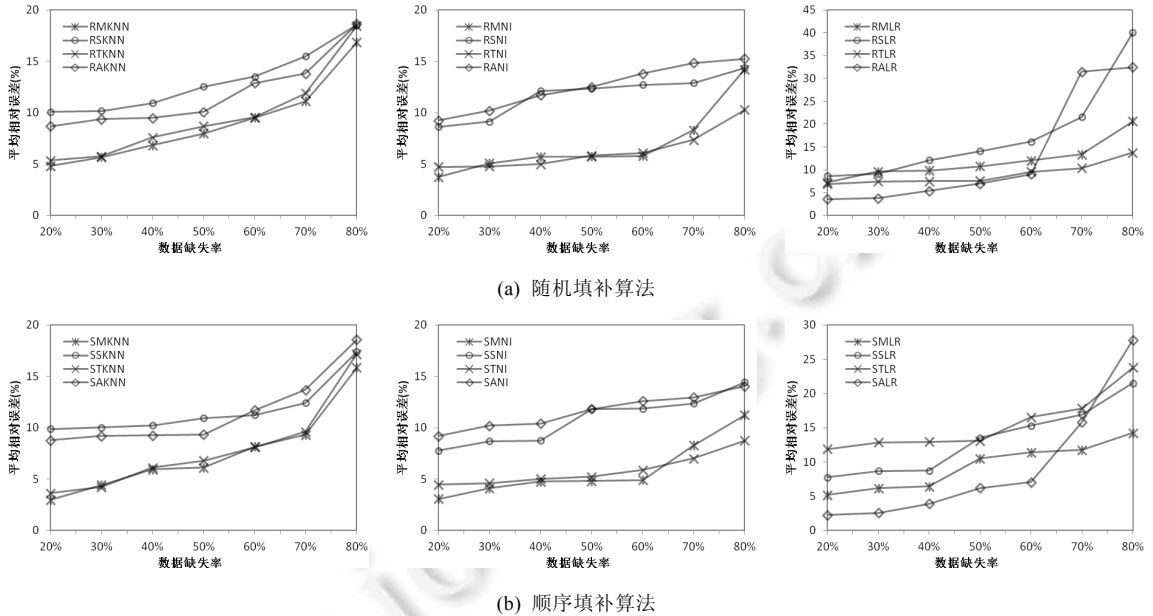


Fig.7 Influence of multimodal vs. single-modal on imputation accuracy (WEATHER)

图 7 多模态 vs.单模态相关性对准确性影响(WEATHER)

从图中可以看出,利用多维度相关性的缺失值填补准确性普遍高于利用单维度相关性的缺失值填补准确性.这是由于在缺失数据较为密集的情况下,利用单维度相关性得到的缺失数据源的近邻节点较少,或在限定近邻点个数(KNN 中  $K$  值)情况下得到的近邻节点与缺失数据源本身的相似度不高,从而导致填补准确性下降.并且从图中可以看出,随着数据缺失率的增加,填补值的平均相对误差相应增大.这是由于随着数据缺失率的增加,可用于填补的完整数据相对减少,即使在顺序填补过程中将已填补值作为观测值用于后续填补,但在实际应用中填补值本身也存在一定误差,填补精度也会相应降低.从图 6 和图 7 可以看出,SENSOR 数据集的平均相对误差较小.这是由于 SENSOR 数据集由 54 个 sensor 对一个房间的温度、湿度进行测量,每个 sensor 的观测值相差都不大,其感知数据的单维度相关性也更强,因此得到的填补结果更为准确.对比之下,WEATHER 数据是美国 30 个不同城市的温度和湿度气象数据,由于各个城市间距离相对较远,且各个城市的感知数据在时间和属性上的波动相对较大,因此即使根据多维度相关性得到的填补值,平均相对误差较大.

5.2 单维度和多维度相关性对填补准确性的影响

本节通过图 8 和图 9 给出了在两个数据集上,基于 KNN,NI,LR 这 3 种填补算法,利用多维度相关性进行顺序填补和随机填补时得到的填补结果准确性的对比.

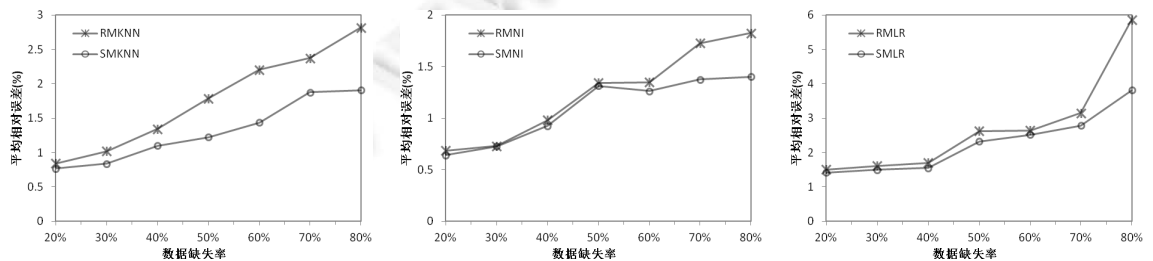


Fig.8 Multimodal sequential vs. random imputation on imputation accuracy (SENSOR)

图 8 多模态顺序 vs.随机填补误差(SENSOR)

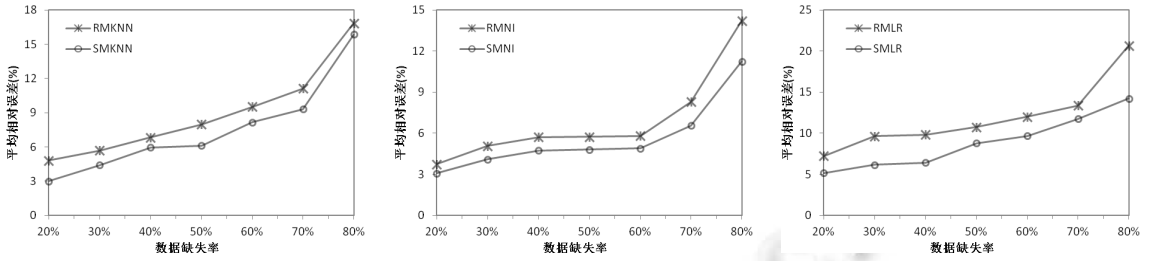


Fig.9 Imputation accuracy of multimodal sequential vs. random imputation (WEATHER)

图 9 多模态顺序 vs.随机填补误差(WEATHER)

从图中可以看出,顺序填补的准确性高于随机填补.这是由于顺序填补考虑了每个缺失数据源近邻的分布,先填补近邻中完整数据相对较多且近似度较高的缺失数据源,由此得到的填补值能更好地为后续缺失值填补所用,得到的填补值准确性也更高.极端情况下,对于近邻节点均为数据缺失的待填补数据源来说,无法利用相关性对其进行填补,只能根据待填补维度的数值变化区间进行随机填补,由此得到的填补值的准确性无疑是很低的.

5.3 各填补算法间的准确性对比

本文在该节给出了基于近邻(NI)的填补算法与基于 KNN(K=5)和多元线性回归(LR)填补算法的准确性比较结果.从图 10 中可以看出,本文提出的 NI 算法准确性最高.这是由于 KNN 填补算法中固定的 K 值无法满足所有缺失数据源的具体情况.对于近邻节点较多的缺失数据源来说,K 值过小会导致由于没有充分利用与待填补数据源相似度较高的近邻节点而引起填补值精度下降;对于近邻节点较少的缺失数据源来说,K 值过大会导致在缺失值估计过程中由于利用了与待填补数据源相似度过低的近邻节点而影响填补精度.因此,固定的 K 值很难同时满足所有缺失数据源的需求.在利用多元线性回归对缺失值进行估计的过程中包含矩阵求逆的运算,而当缺失数据源周围的近邻节点较多时,根据近邻节点及历史数据组成的求逆矩阵很容易是奇异矩阵,因而导致回归模型建立失败.因此,基于 LR 的填补算法在 3 种填补算法中准确性最低.

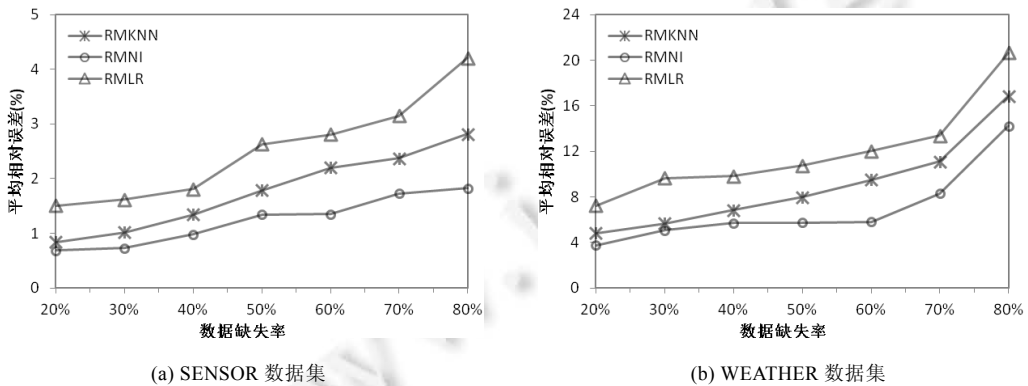


Fig.10 Accuracy of KNN, NI, LR based imputation

图 10 KNN 填补、NI 填补、LR 填补的准确性

5.4 各填补算法的执行时间

最后,本文给出了基于顺序的 3 种填补算法的执行时间随着数据量以及数据缺失率增长的变化情况.图 11 给出的是 3 种算法的执行时间随数据量的变化,其中,纵坐标为对数刻度,数据缺失率设定为 50%.从图中可以看出,算法的执行时间随着数据量的增加呈线性增长.这是由于随着数据量的成倍增加,缺失数据源的个数也成倍

增加.图 12 给出的是随着数据缺失率的增加,3 种基于顺序的填补算法的执行时间,其数据量设定为 10K.从图中可以看出:随着数据缺失率的增加,3 种缺失值填补算法的执行时间也随着增加,并且基于回归的填补算法执行时间最长,基于 NI 的填补算法执行时间最短.这是由于基于回归的填补算法需根据历史数据对线性填补模型的参数进行学习,而 NI 填补算法只需根据相似度阈值对近邻节点进行筛选.此外,通过图 11 和图 12 的实验结果可以看出:当数据量较大时,算法的执行时间较长.例如:当数据缺失率为 50%、数据量为 10M(43 万条数据)的情况下,其执行时间大约需要 10min.因此,在对感知数据流进行实时填补过程中,如何设计有效的增量填补算法以提高执行效率,是一个很有挑战的问题,我们将在未来的工作中对其进行研究.

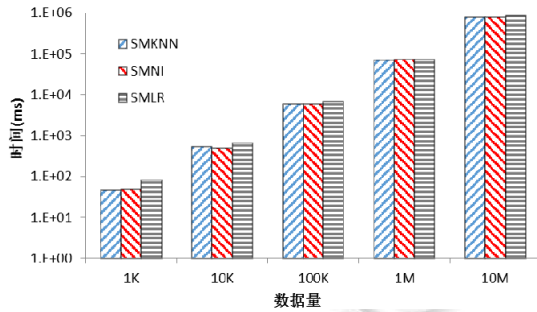


Fig.11 Run time vs. data size  
图 11 执行时间 vs.数据量

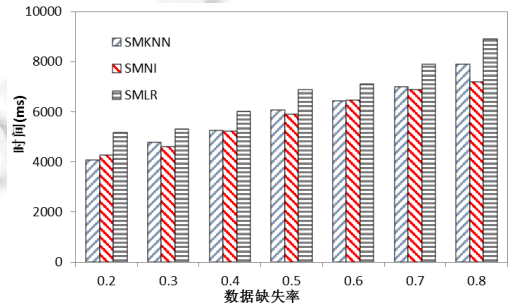


Fig.12 Run time vs. missing ratio  
图 12 执行时间 vs.缺失率

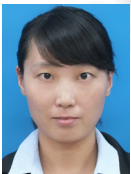
## References:

- [1] Racine J, Li Q. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 2004,119(1):99–130. [doi: 10.1016/S0304-4076(03)00157-X]
- [2] Zhu XF, Zhang SC, Jin Z, Zhang ZL, Xu ZM. Missing value estimation for mixed-attribute data sets. *IEEE Trans. on Knowledge and Data Engineering*, 2011,23(1):110–121. [doi: 10.1109/TKDE.2010.99]
- [3] Zhou X, Wang X, Dougherty ER. Missing-Value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics*, 2003,19(17):2302–2307. [doi: 10.1093/bioinformatics/btg323]
- [4] Qin YS, Zhang SC, Zhu XF, Zhang JL, Zhang CQ. POP algorithm: Kernel-Based imputation to treat missing values in knowledge discovery from databases. *Expert Systems with Applications*, 2009,36(2):2794–2804. [doi: 10.1016/j.eswa.2008.01.059]
- [5] Velicer WF, Colby SM. A comparison of missing-data procedures for ARIMA time-series analysis. *Educational and Psychological Measurement*, 2005,65(4):596–615. [doi: 10.1177/0013164404272502]
- [6] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 2001,17(6): 520–525. [doi: 10.1093/bioinformatics/17.6.520]
- [7] Joensuu DW, Bankhofer U. Hot deck methods for imputing missing data. In: *Proc. of the Machine Learning and Data Mining in Pattern Recognition*. Berlin, Heidelberg: Springer-Verlag, 2012. 63–75. [doi: 10.1007/978-3-642-31537-4\_6]
- [8] David I, Michael PB, Abt A. Weighted sequential hot deck imputation: SAS Macro vs. SUDAAN's PROC HOTDECK. In: *Proc. of the SAS Global Forum*. 2013. 213–2013.
- [9] Zhang CQ, Zhu XF, Zhang JL, Qin YS, Zhang SC. GBKII: An imputation method for missing values. In: *Proc. of the Advances in Knowledge Discovery and Data Mining*. 2007. 1080–1087. [doi: 10.1007/978-3-540-71701-0\_122]
- [10] Zhang S. Parimputation: From imputation and null-imputation to partially imputation. *IEEE Intelligent Informatics Bulletin*, 2008, 9(1):32–38.
- [11] Caruana R. A non-parametric EM-style algorithm for imputing missing values. In: *Proc. of the Artificial Intelligence and Statistics*. 2001.
- [12] Meng XL, Rubin DB. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 1992,79(1):103–111. [doi: 10.1093/biomet/79.1.103]

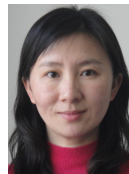
- [13] Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 2001,27(1):85–96.
- [14] Aittokallio T. Dealing with missing values in large-scale studies: Microarray data imputation and beyond. *Briefings in Bioinformatics*, 2010,11(2):253–264. [doi: 10.1093/bib/bbp059]
- [15] Mihail H, Gruenwald L. Estimating missing values in related sensor data streams. In: *Proc. of the COMAD*. 2005. 83–94.
- [16] Jiang N, Gruenwald L. Estimating missing data in data streams. In: *Proc. of the Advances in Databases: Concepts, Systems and Applications*. Berlin, Heidelberg: Springer-Verlag, 2007. 981–987. [doi: 10.1007/978-3-540-71703-4\_89]
- [17] Christos A, Peter T. Scaling out big data missing values imputations. In: *Proc. of the SIGKDD*. 2014. 651–660. [doi: 10.1145/2623330.2623615]
- [18] Zheng Y, Liu F, Hsieh HP. U-Air: When urban air quality inference meets big data. In: *Proc. of the SIGKDD*. 2013. 1436–1444. [doi: 10.1145/2487575.2488188]
- [19] Kim KY, Kim BJ, Yi GS. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics*, 2004, 5(1):159–167. [doi: 10.1186/1471-2105-5-159]
- [20] Verboven S, Branden KV, Goos P. Sequential imputation for missing values. *Computational Biology and Chemistry*, 2007,31(5): 320–327. [doi: 10.1016/j.compbiolchem.2007.07.001]
- [21] Pan LQ, Li JZ, Lao JZ. A temporal and spatial correlation based missing values imputation algorithm in wireless sensor networks. *Chinese Journal of Computers*, 2010,33(1):1–11 (in Chinese with English abstract). <http://cjc.ict.ac.cn/qwjs/view.asp?id=3008>

#### 附中文参考文献:

- [21] 潘立强,李建中,骆吉洲. 传感器网络中一种基于时-空相关性的缺失值估计算法. *计算机学报*, 2010,33(1):1–11. <http://cjc.ict.ac.cn/qwjs/view.asp?id=3008>



马茜(1988—),女,河北秦皇岛人,硕士生,CCF 学生会员,主要研究领域为感知数据管理.



李芳芳(1977—),女,博士,讲师,CCF 高级会员,主要研究领域为数据库技术,传感器网络 CPS 数据管理.



谷峪(1981—),男,博士,副教授,CCF 高级会员,主要研究领域为图,空间数据管理.



于戈(1962—),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据管理理论与技术,分布与并行系统.