DOI: 10.7641/CTA.2015.40828

基于最大方差权信息系数的煤气数据填补

吕 政, 赵 珺†, 刘 颖, 王 伟

(大连理工大学 控制科学与工程学院, 辽宁 大连 116033)

摘要: 在数据的挖掘、建模与优化领域, 数据的完整性与准确性是进行此类研究的基础. 鉴于冶金能源系统的复杂性和现场数据采集过程易受干扰的特点, 其数据在获取过程中极易发生数据缺失的现象, 从而造成模型无法建立, 隐含信息无法准确挖掘等情况. 本文针对钢铁企业副产煤气的发生、消耗流量数据出现的缺失情况, 通过分析相似工况下能源流量数据的相关特性, 提出一种基于最大方差权信息系数的冶金企业副产煤气系统流量数据填补方法. 该方法针对现场经常发生的两类数据缺失情况, 即数据点间断缺失和数据长时间连续缺失, 以最大方差权信息系数作为样本筛选准则, 并采用基于核学习的方法对缺失数据进行填补. 为验证本文提出的数据填补方法的有效性, 本文对上海宝钢高炉、焦炉和冷热轧用户的实际生产数据的运行试验, 结果表明该方法相比其他的方法在填补精度上有很大优势.

关键词: 冶金能源系统; 数据填补; 样本筛选; 最大方差权信息系数

中图分类号: TP206+.3　　**文献标识码**: A

Missing data imputation based on maximal variance weight information coefficient for gas flow in steel industry

LÜ Zheng, ZHAO Jun†, LIU Ying, WANG Wei

(School of Control Science and Engineering, Dalian University of Technology, Dalian Liaoning 116033, China)

Abstract: In data-driven-based modeling and optimization, the completeness and the accuracy of data are the foundations for further research tasks. Since the energy system of steel industry is rather complicate and its data-acquisition process might be frequently affected by the malfunctions of data transportation, storage and transformation, the data-missing phenomenon usually occurs, which might lead to the failure of model building or accurate information discovery. In this study, by analyzing the correlation of the energy data with respect to corresponding operation conditions in manufacturing, a data imputation method for the missing data in the byproduct gas flow is proposed. In this method, the proposed maximal variance weight information coefficient (MVWIC) is adopted as the sample selection criteria to realize the data imputation by using the kernel-learning-based method. To validate the proposed method, two types of missing modes that frequently occurred in steel industry are considered here, i.e., the intermittent missing and the long-term continuous missing. A series of experiments using the practical energy data indicates that the proposed method exhibits good performance on the imputation accuracy when compared to other methods.

Key words: energy system of steel industry; data imputation; sample selection; maximal variance weight information coefficient

1 引言(Introduction)

随着现代工业信息化水平的逐步提高, 工业企业在生产过程中积累了大量的实时数据. 深入挖掘隐含在此类数据背后的信息, 并将其应用在生产自动化与智能化的解决方案中, 是当前工业信息化领域的研究热点[1]. 在冶金工业, 副产煤气是炼焦、加热炉、电厂等环节必需的二次能源, 其有效合理的利用直接影响到钢铁企业的能耗标准和产出成本[2], 因此生产现场对副产煤气流量的实时波动情况相当重视[3]. 基于数据和知识挖掘的方法已经成为该领域解决实际问题的可行方法[4–6]. 然而此类方法的重要前提是现场数据获取的完整性和准确性. 鉴于冶金工业现场情况复

收稿日期: 2014−09−06; 录用日期: 2015−01−06.

†通信作者. E-mail: zhaoj@dlut.edu.cn; Tel.: +86 411-84707582.

国家“863”计划项目(2013AA040703), 国家自然科学基金项目(61034003, 61304213, 61104157, 61273037, 61473056), 中央高校基本科研业务费专项资金项目(DUT13RC203)资助.

Supported by National High-Tech Research & Development Program (2013AA040703), National Natural Science Foundation of China (61034003, 61304213, 61104157, 61273037, 61473056) and Fundamental Research Funds for the Central Universities (DUT13RC203).

杂, 采集器故障、传输偏差、存储失误等情况时有发生. 企业能源中心获得的实时数据极易出现缺失的情况, 且其缺失往往无固定的规律可循, 属于完全随机缺失机制(missing completely at random, MCAR)[7]. 为此, 采用合理的方式完成对缺失数据的填补工作就显得尤为重要, 目前数据填补方法的研究逐渐被诸多学者所重视.

在传统数据填补研究领域, 常用的方法有均值填补法、期望值最大化法、回归填补法等等[8–10]. 文献[11]回顾了不同形式的热平台方法以及目前对其统计特性的研究. Schafer和Maren提出了多变量缺失数据的多重替代法[12]. James对EM算法进行改进, 提出Monte Carlo EM算法, 并在三类数据集上进行了验证, 但是对于算法收敛速度慢的问题没有得到很好的解决[13]. 文献[14]采用MLR(multinomial logistic regression)方法估计软件成本数据库中缺失的数据. 这些方法虽然计算简单, 但是当数据波动较大、数据缺失点增多时, 填补的精度将大大下降.

近几年国内外学者在保留传统方法研究的基础上融入了机器学习、统计分析等方法[15–17]. 文献[18]提出一种基于缺失值分析的自回归模型, 该模型适用于某一个特定时间点包含大量缺失数据或者数据全部缺失的情况. 文献[19]中, Buure在特定条件下使用多重填补法同时对离散型和连续型缺失值进行填补, 并将结果与文献[20]的经典结果进行比较, 提出一种多重填补法的改进方案. 以上这些方法较传统的方法在填补精度上有很大的提高, 并能够较好的解决数据大量缺失的情况, 但是其研究对象都是针对具有多属性的数据, 即每条数据均由多个相关的变量组成, 因此对于时间序列数据, 上述方法将受到限制. 目前, 针对时间序列缺失数据填补方法的国内外文献相对较少. 其中文献[21]对比了处理时间序列数据缺失的4种方法: 缺值删除、平均替代、相邻均值、极大似然估计, 在不同的数据缺失程度下的填补效果, 但是该实验所研究的时间序列数据具有标准的正态分布特性, 且其模型是可被准确估计, 而对于冶金企业的副产煤气数据很难具有此类特性.

考虑冶金能源系统数据的时间序列特性及能源用户在类似工况下煤气流量数据的相关性特点, 本文提出一种基于最大方差权信息系数的冶金煤气流量数据填补方法. 该方法以最大方差权信息系数作为评判准则, 通过分析不同时间段流量数据间的相关性寻找与待填补数据处于相似工况的样本数据, 进而用核学习的方法建立数据缺失点的学习模型, 实现数据填补. 根据能源中心现场的实际情况, 本文将冶金能源系统的数据缺失问题归结为两大类, 数据间断性缺失和数据连续大量缺失. 基于最大方差权信息系数的填补方法可以很好地解决这两类问题, 实际数据运行结果表明, 相比于其他方法, 该方法在填补精度上有很大提高.

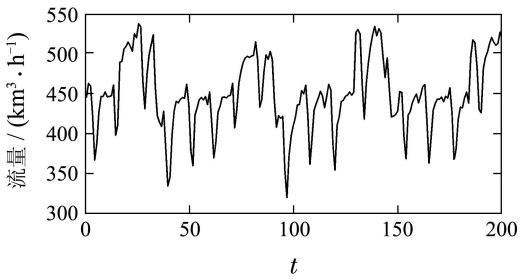
本文具体安排如下: 第2节简要描述冶金能源系统中存在的数据缺失问题及其数据间存在的关联特性. 第3节给出了最大方差权信息系数的概念, 并针对数据间断性缺失和连续大量缺失两种情况, 阐述了将这一方法应用于缺失数据填补问题. 第4节通过对比实验说明本文所提出的填补方法对冶金煤气流量缺失数据填补的有效性. 第5节是对全文的总结.

2　问题描述及数据特征(Problem description and data characteristics)

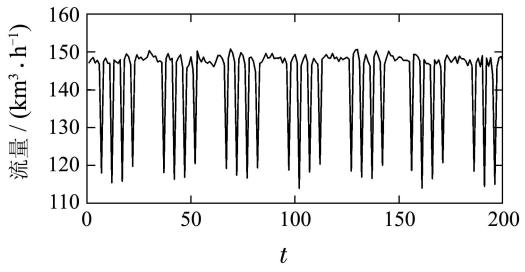
钢铁工业是能源消耗大户, 投入的一次能源有大约40%转化成为副产煤气. 因此, 对副产煤气的发生量和使用量进行有效的监控和分析, 对于钢铁企业降低成本、节约能源具有重要意义. 目前, 世界上各大型冶金企业已建立了遍布全厂的SCADA(supervisory control and data acquisition)系统, 可以对多种能源介质进行实时监控. 然而由于冶金工业现场情况复杂, 采集器故障、传输偏差、存储失误等原因常常导致采集数据的缺失, 给现场人员的生产过程分析和平衡调度决策工作带来了困难.

目前, 在实际应用中如果出现数据缺失, 多采用人工估计的方法, 根据生产经验对缺失数据进行估计. 然而由于调度人员的经验差异以及设备工况的变化等问题, 这种方法实现起来难度大, 其精度也无法满足数据填补的要求. 特别是针对缺失数据较多的情况, 更是如此.

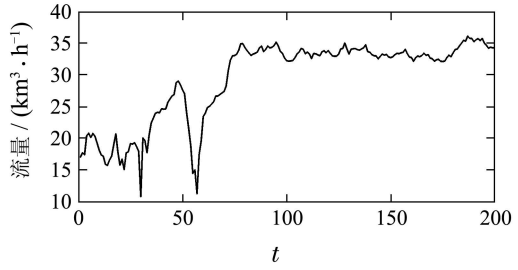
从冶金能源系统的数据特征来看, 不同的副产煤气发生单元及其用户所涉及的煤气流量数据特征各有不同. 其特征依据大致情况可归结为以下几类: 1) 类似周期性数据, 如高炉用户, 其生产计划较为恒定, 但生产过程不稳定, 受环境因素影响较大, 从而导致煤气流量在周期性和幅值上都会存在一定的变化(图1(a)). 2) 存在高频特性的数据, 如焦炉用户, 其生产情况稳定, 但存在剧烈的波动, 具有明显的周期性(图1(b)). 3)无规律波动数据, 如冷热轧用户, 其煤气使用流量受加工原料和生产方式等因素影响较大, 因此呈现出不规律的变化(图1(c)).



(a) 4#高炉煤气(blast furnace gas, BFG)发生流量



(b) 1.2#焦炉BFG使用流量



(c) 1800冷轧BFG使用流量

图 1 冶金能源煤气流量示意图

Fig. 1 Gas flow diagram of steel industry

3 基于最大方差权信息系数的数据填补 (Data imputation based on maximal variance weighted information coefficient)

通过现场数据特征分析, 笔者发现煤气数据的变化主要是由于生产计划以及加工产品的不同所引起的, 而同一生产设备在工况相似的情况下, 其煤气流量具有一定的相关特性. 鉴于此情况, 本文提出一种基于数据相关性分析的煤气流量数据填补方法, 通过计算最大方差权信息系数分析数据段间的相关性, 找到相似工况下的生产数据, 并结合核学习算法对筛选的数据进行建模, 进而实现对缺失数据的填补.

3.1 最大信息系数(Maximal information coefficient)

目前针对数据相关性的量化分析方法, 如欧氏距离[22]、马氏距离[23]、相关系数[24]等等, 都是计算数据的相似度或线性相关性, 对于非线性的复杂关系则不适用. 最大信息系数(maximal information coefficient, MIC)是在两个变量的散点图上用网格将点分割, 通过计算在不同网格维数下的最大互信息量得到的, 可以准确地衡量变量间的相关性[25]. 假设两事件X和Y, 其互信息[26]可定义为

$$I(X,Y)=H(X)+H(Y)-H(X,Y), \tag{1}$$

其中$H(X,Y)$是联合熵, 即

$$H(X,Y)=-\sum p(X,Y)\log p(X,Y), \tag{2}$$

$p(X,Y)$是联合概率. 对于一个有限的数据集$D\{x, y\}$, 将x值划分a个区间, 将y值划分b个区间, 即得到在数据集D上的$b\times a$维网格G. 记此划分为分布$D|_G$, 则变量x, y的最大互信息$I^*(D,a,b)$为

$$I^*(D,a,b)=\max I(D|_G), \tag{3}$$

其中$I(D|_G)$表示分布$D|_G$的互信息. 互信息最大值的搜索范围是所有a列b行的网格划分, 如图2所示, 并通过归一化处理, 得到特征矩阵$M(D)$:

$$M(D)_{a,b}=\frac{I^*(D,a,b)}{\log\min\{a,b\}}. \tag{4}$$

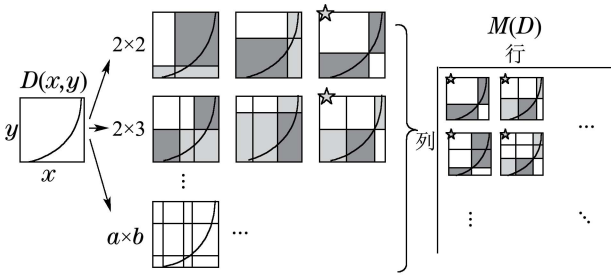


图 2 特征矩阵的生成

Fig. 2 Generation of the characteristic matrix

那么对于有n条记录的数据集D, 其MIC定义为

$$\mathrm{MIC}(D)=\max_{ab<B(n)}\{M(D)_{a,b}\}, \tag{5}$$

其中

$$\omega(1)<B(n)<O(n^{1-\varepsilon}),\ 0<\varepsilon<1. \tag{6}$$

MIC在分析两个变量的相关性时具有很好的效果, 但是要准确的判断出两个变量的相关性, 需要具有足够的数据样本, 在数据样本量较少的情况下, 其准确度将大大降低[25]. 而在冶金煤气流量数据填补的问题中, 由于不同用户的生产特性以及核学习方法对数据输入维数的限制, 其相关性的评价长度较短, 为了适应少量数据分析的情况, 同时考虑到工业采集数据含噪高的特点, 本文提出一种最大方差权信息系数, 在MIC基础上加入方差信息, 从而提高冶金煤气数据的相关性分析效果.

3.2 最大方差权信息系数 (Maximal variance weighted information coefficient)

本文提出一种最大方差权信息系数(maximal variance weighted information coefficient, MVWIC), 可在冶金工业样本数据少的情况下对其煤气流量数据的相关性进行较好的分析, 从而有效地判断工况的相似性. 计算MIC的过程核心在于特征矩阵$M(D)$的生成, 即在不同的网格划分下数据集D的最大互信息, 如下所示:

$$\begin{aligned}&\max\{I(D,a,b)\}=\\&\max\{H(P)+H(Q)-H(P,Q)\},\end{aligned} \tag{7}$$

其中: P为数据集D的列划分, Q为行划分, a为网格

的列数, b为网格的行数. 式(7)是一个具有两维度的极值问题, 因此在求解时先将数据集D均分为b行, 即首先确定Q, 然后用动态规划的方法对列分布P进行优化求解, 从而进一步得到如式(8)所示的优化目标.

$$\begin{aligned}&\max\{H(P)-H(P,Q)\}=\\&\max\{\sum_{j=1}^{b}\frac{\#_{*,j}}{m}\log\frac{m}{\#_{*,j}}-\sum_{j=1}^{b}\sum_{i=1}^{|Q|}\frac{\#_{i,j}}{m}\log\frac{m}{\#_{i,j}}\}=\\&\max\{\sum_{j=1}^{b}\sum_{i=1}^{|Q|}\frac{\#_{i,j}}{m}\log\frac{\#_{i,j}}{\#_{*,j}}\},\end{aligned}\tag{8}$$

其中: m为数据集D中数据点的总个数, $\#_{*,j}$表示D中的数据点落在网格G第j列的个数, $\#_{i,j}$表示D中的数据点落在网格G第i行第j列的个数最优的列划分就是尽可能的让数据点聚集在一起. 这样求得的信息系数最大. 如果在数据量比较集中的列, 其数据波动超出平均水平, 也就意味着该数据集的相关性较小. 由于方差可以描述数据的离散程度, 因此在相关性较小的情况下, $\mathrm{var}_{*,j}>\overline{\mathrm{var}}_{*,j}$, 即$\dfrac{\mathrm{var}_{*,j}}{\overline{\mathrm{var}}_{*,j}}>1$. 反之, 如果多数数据都集中在波动比较平稳的列上, 说明数据集的相关性比较大, 而此时$\dfrac{\mathrm{var}_{*,j}}{\overline{\mathrm{var}}_{*,j}}<1$. 由于式(8)计算值为负数, 为了使计算的信息系数在小样本的情况下克服噪声的影响加快收敛, 本文将方差信息加入到相关性分析的过程中, 在其目标函数中引入方差权重, 进一步得到如式(9)所示的优化目标, 即MVWIC.

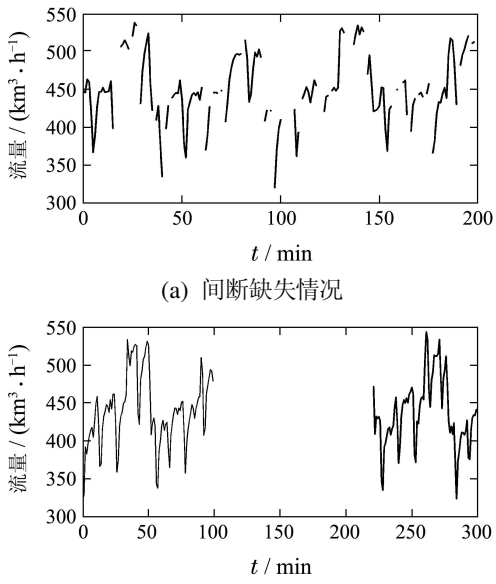
$$\max\{\sum_{j=1}^{b}\sum_{i=1}^{|Q|}\frac{\mathrm{var}_{*,j}}{\overline{\mathrm{var}}_{*,j}}\frac{\#_{i,j}}{m}\log\frac{\#_{i,j}}{\#_{*,j}}\},\tag{9}$$

其中: $\mathrm{var}_{*,j}$为网格第j列中y值的方差, $\overline{\mathrm{var}}_{*,j}$为各列$y$值方差的平均值.

可以看出, 对于MVWIC, 当数据相关性较小时, 计算的相关性系数将更加接近于0, 而在相关性较大时, 计算结果更加接近于1, 从而加强了对冶金煤气流量数据的相关性评判能力.

3.3 基于信息系数的数据填补(Data imputation based on information coefficient)

本文所研究的煤气流量的数据缺失问题属于完全随机缺失机制. 从采集到的历史数据中笔者发现, 数据的缺失可以分为两种情况, 一种是间断性缺失, 数据缺失位置比较分散、随机, 一般连续缺失点的个数不会太多. 另一种情况是数据连续性缺失, 这种情况下用户在较长的一段时间内都没有采集值, 连续缺失点较多, 一般是由于采集器故障或设备检修导致. 两种缺失情况如图3所示, 通过采用本文所提的MVWIC可以有效分析煤气流量数据的相关性, 从中找出具有相似工况特征的数据, 进而为数据填补奠定基础.



(a) 间断缺失情况

(b) 连续缺失情况

图 3 4#高炉BFG发生流量数据缺失情况

Fig. 3 Data missing for BFG generation of #4 blast furnace

情况 1 数据间断缺失.

数据的间断性缺失主要有两个原因. 一种是由于工业生产现场复杂产生的一些偶然情况的干扰, 例如数据通讯或采集器受到扰乱, 一般这种缺失多为短时期的, 因此本文可以从该用户的历史数据中找到足够长的连续数据用来建模分析. 另一种情况是由于某些设备的老化使得采集到的数据长久以来都存在间断性缺失的问题. 因为煤气流量数据为时间序列形式, 因此在进行相似性对比之前, 本文先要对数据进行分析, 根据用户的实际生产特性和历史数据的分析确定其最优评价长度m. 由于过高的数据缺失问题会影响相似性分析和核学习方法的效果, 因此本文对数据缺失点采取逐一填补的策略, 每次针对一个缺失点进行填补, 并循环此操作, 具体步骤如下:

1) 数据准备.

假设有数据点$x_1,x_2,\cdots,x_n$, 本文首先选择包含最少缺失点的数据段$x_i,x_{i+1},\cdots,x_{i+m-1}$作为数据填补的目标. 若$x_{i+j}$为本次待填补的数据点, 那么对于其他缺失点暂用相邻数据的均值替代.

2) 寻找相似工况数据.

用MVWIC判断$x_i,x_{i+1},\cdots,x_{i+m-1}$与其附近连续数据样本的相关性, 并将大于MVWIC阈值的数据段存储到训练样本集中. 根据各个用户不同的数据情况, 本文将设定的训练样本的长度作为寻找相似工况数据的终止条件.

3) 建立数据填补模型.

将训练样本集合中x_{i+j}位置的数据作为模型输出, 其他位置的数据作为输入, 用最小二乘支持向量机(LSSVM)训练得到关于缺失点的填补模型. 最后将$x_i,\cdots,x_{i+j-1},x_{i+j+1},\cdots,x_{i+m-1}$代入该模型, 即

可得到x_{i+j}的填补值.

情况 2 数据连续缺失.

在生产过程中, 当发生通讯或者采集器故障时, 现场实时监控数据会无法传送到数据库. 这时, 需要工作人员到达现场进行故障排除, 由于要解决这种问题难免要花费一些时间, 从而造成了一段时间内的生产数据缺失. 缺失数据的长度取决于排除故障所用的时间, 如果问题严重可能会缺失几小时甚至半天的数据. 对于这种连续大量的数据缺失问题, 本文同样首先确定其最优评价长度m. 同时, 考虑到填补误差的累积以及数据段前后工况的匹配, 本文对缺失点采用从两边到中间的填补顺序, 每次针对一个缺失点进行填补, 并循环次操作, 具体步骤如下:

1) 寻找相似工况数据.

假设数据连续缺失点为$x_i, x_{i+1}, \cdots, x_j$, 连续缺失数据段前后的采集数据为完整数据. 若$x_i$为待填补点, 选取$x_{i-m}, \cdots, x_{i-1}$作为相关性比较的目标. 从历史数据中选取连续$m$时间长度的样本数据, 用MVWIC判断两段数据的相关性, 并将大于MVWIC阈值的数据段存储到训练样本集中. 本文同样用训练样本的长度作为寻找相似工况数据的终止条件.

2) 建立数据填补模型.

以填补x_i为例, 将训练样本中x_i对应位置的数值作为输出, 其他数据作为输入, 用最小二乘支持向量机(LSSVM)训练得到关于缺失点的填补模型. 最后将$x_{i-m}, \cdots, x_{i-1}$代入该模型即可得到x_i的填补值.

4 实验分析(Experiments and analysis)

本文的仿真实验所用到的数据来自于上海宝钢能源中心现场数据库2013年8月的数据, 数据时间粒度为1 min. 为验证本文所提MVWIC的鲁棒性, 本文首先选取具有相似工况的数据段, 用MIC和MVWIC分别进行相关性分析. 在实验中, 本文对原始数据添加不同程度的均匀噪声进行分析, 噪声的强度通过加噪声后的数据与原始数据的相关性进行衡量, 公式如下:

$$\text{噪声强度} = 1 - \frac{(\sum_{i=1}^{n}(x_i-\bar{x})(\acute{x}_i-\bar{\acute{x}}))^2}{\sum_{i=1}^{n}((x_i-\bar{x})^2)\sum_{i=1}^{n}((\acute{x}_i-\bar{\acute{x}})^2)}, \quad (10)$$

其中: x_i为原始数据点, $\acute{x}_i$为加入人工噪声的数据点, $\bar{x}$为原始数据的平均值, $\bar{\acute{x}}_i$为含噪声数据的平均值, n为数据段的长度. 相关性分析结果如图4所示, 图中每种图形表示不同的煤气流量数据. 可以看出, 两种方法得到的结果均随着噪声强度增加而减小, 在噪声强度很低的时候, 用本文方法可以更加有效的判断其相关性. 当噪声强度达到0.2左右时, 从各个用户的煤气流量数据仍能够看出较大的相关性, 具有很好的鲁棒性. 并且随着噪声强度的增加, 其结果相比于原始的MIC数据波动范围较小, 可以更加准确的判断煤气流量数据的相关性.

在实验部分, 为了方便计算数据填补的精度, 选择连续的数据进行实验, 并人工设定数据缺失的位置. 数据缺失率即为缺失点个数与数据段总长度的比值. 以下本文分别针对间断性缺失和连续缺失两种情况与其他数据填补方法进行对比. 本文选取归一化均方根误差(normalized root meat square error, NRMSE)作为评价指标, 具体公式如下:

$$\text{NRMSE} = \sqrt{\frac{1}{n\|Y\|^2}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}, \quad (11)$$

其中: n为填补点个数, $\hat{Y}_i$为填补值, Y为实际值.

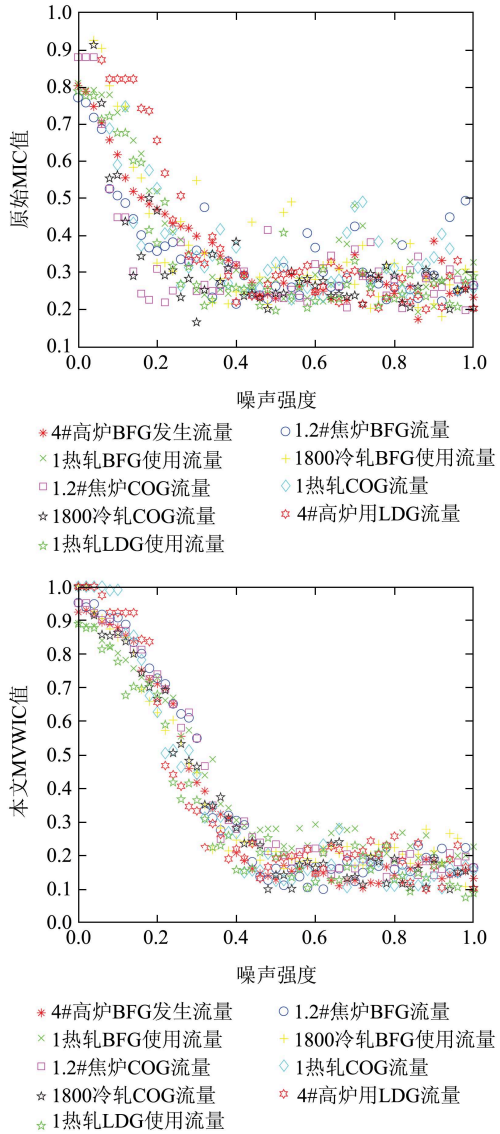


图 4 相同工况下煤气流量数据的相关性分析

Fig. 4 Correlation analysis of gas flow data under the same operation condition

4.1 参数选取(Parameter selection)

1) 最优评价长度.

在通过MVWIC选择具有相关性的数据段之前, 本文首先要确定选取多长的数据进行评价是最优的. 在

实际情况中, 由于生产情况所限, 本文所研究的煤气流量数据只在短时间内存在相关性, 如果时间长度过长, 不同数据段数据的相关性分析将失去意义. 同时, 在核学习方法中, 数据的长度也要有一定的要求, 所选取的数据长度过长或过短都会影响学习效果. 本文根据用户的实际生产情况和历史数据的分析, 考虑现场的工艺特点, 并结合文献[27]的方法确定最优评价长度. 即先根据现场工艺和历史数据的特点确定评价长度的适当范围, 并在该范围内对序列进行相空间重构. 用Gamma Test计算其输入输出数据间的有效噪声, 选取有效噪声最小的输入输出模型, 即为最优的评价长度. 各用户最优评价长度选取如表1所示.

表 1 各煤气用户参数选取

Table 1 Parameter selection

用户名称	最优评价长度	相关性阈值
4#高炉BFG发生流量/($km^{-3}\cdot h^{-1}$)	75	0.75
1.2#焦炉BFG流量/($km^{-3}\cdot h^{-1}$)	40	0.90
1热轧BFG使用流量/($km^{-3}\cdot h^{-1}$)	36	0.80
1800冷轧BFG使用流量/($km^{-3}\cdot h^{-1}$)	38	0.80
4#高炉用LDG流量/($km^{-3}\cdot h^{-1}$)	70	0.75
1热轧LDG使用流量/($km^{-3}\cdot h^{-1}$)	75	0.80
1.2#焦炉COG流量/($km^{-3}\cdot h^{-1}$)	40	0.90
1热轧COG流量/($km^{-3}\cdot h^{-1}$)	36	0.85
1800冷轧COG使用流量/($km^{-3}\cdot h^{-1}$)	38	0.85

2) $B(n)$.

$B(n)$即MVWIC计算过程中网格的大小限制, $B(n)$越大, 划分数据的行列数越大. 因此, 如果$B(n)$过大, 即使对于随机的变量, 其每一个数据点都可能会有其独立的网格单元, 计算的互信息量就会很大. 而当$B(n)$过小时, 通过MVWIC只能寻找到比较简单的关系, 对于相对复杂的关系, 从MVWIC值无法体现出来. 如文献[25]所述, 当$B(n)$的增长速度比n慢时, 可以很好的协调以上的问题, 因此本文选取$B(n) = n^{0.7}$.

3) 相关性阈值.

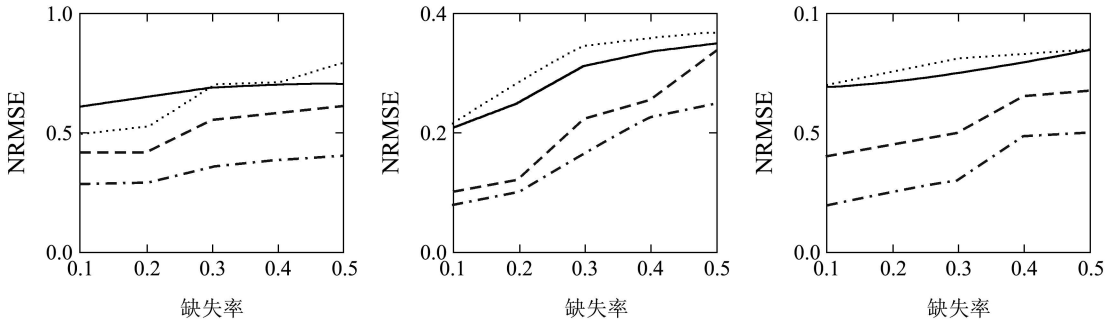
由于噪声和评价长度的限制, 计算得到的MVWIC值很难收敛到1或者0, 因此本文引入了MVWIC阈值来决定数据的相关性. 不同用户的阈值如表1所示.

4.2 实验结果(Experiment results)

情况 1 数据间断缺失.

本文从能源系统中选取高炉、焦炉、冷轧、热轧几个典型用户对BFG, 焦炉煤气(coke oven gas, COG), 转炉煤气(linz-donawitz gas, LDG)3种煤气的发生和使用量作为研究对象, 从各用户历史数据中选取2013年8月30日的数据, 用生成随机数的方式决定这段数据的缺失点位置. 考虑数据缺失率由10%变化至50%的情况下, 不同方法填补效果的对比. 本文采用的是基于MVWIC判断数据相关性进行回归建模的方法, 因此这里选取基于欧氏距离判断数据相关性进行回归建模的方法, 时间序列预测方法, 以及传统的样条插值方法与本文方法进行对比. 其中, 时间序列预测方法要求训练样本数据连续, 因此本文选取足够长的连续数据为其训练模型, 其他3种方法对数据连续性无特殊要求. 不同方法的数据填补效果如图5和表2所示.

从实验可以看出, 对于冶金煤气流量数据来说, 采用通过相似工况下的生产数据进行建模的方法, 其精度明显好于其他方法. 在数据缺失率从10%升至50%的情况下, 该方法的补点精度略有下降, 但是足够满足现场的要求. 尤其对于工况变化较大的冷热轧用户, 通过分析相似工况下的数据特点进行数据填补, 相比于其他的数据填补方法具有更高的精度. 采用本文方法进行工况的相似性判断可以充分的考虑生产原料、环境噪声等各类因素的干扰, 对于冶金煤气流量数据具有更好的适用性, 同时, 该方法对于历史数据的连续性要求不高, 因此对于实际应用具有很大意义.



(a) 4高炉BFG受入流量　(b) 1.2#焦炉BFG　(c) 1热轧BFG

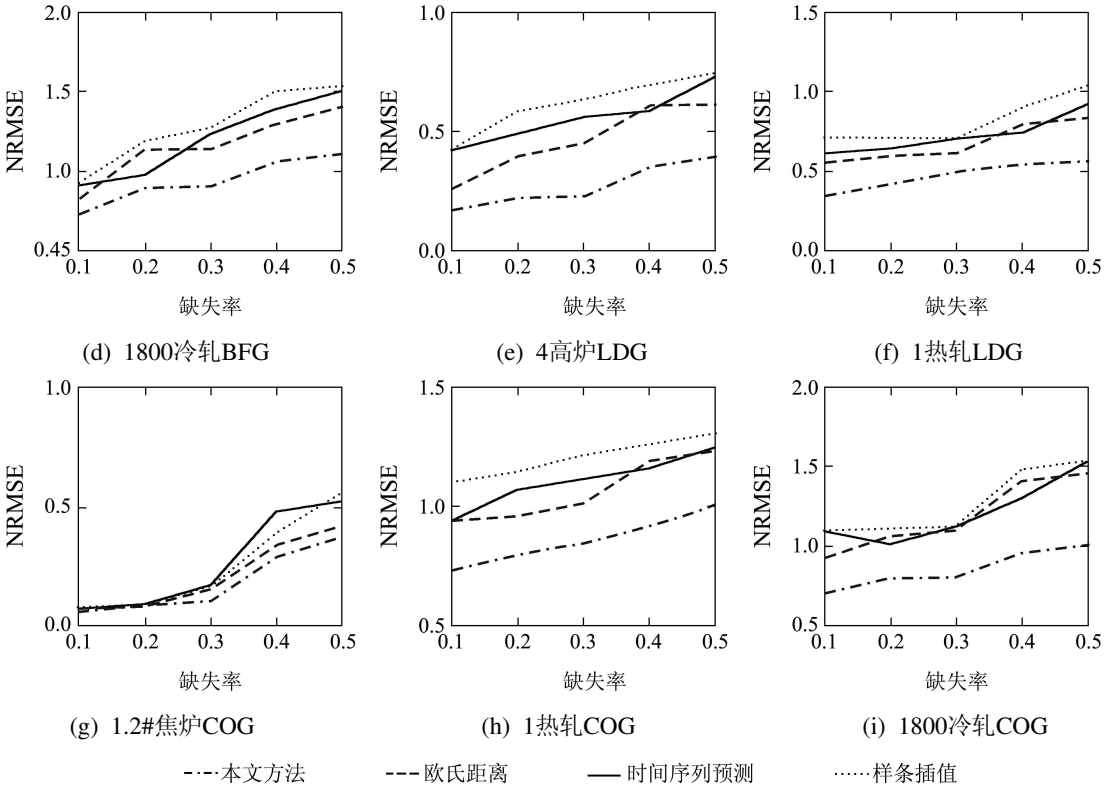


图 5 间断缺失数据的填补精度

Fig. 5 Imputation accuracy for Intermittent missing data

表 2 不同缺失率下的平均填补精度NRMSE

Table 2 Average NRMSE of different missing rates

用户名称	本文方法	欧式距离	时间序列预测	样条插值
4#高炉BFG发生流量/($km^{-3}\cdot h^{-1}$)	0.3465	0.5160	0.6710	0.6445
1.2#焦炉BFG流量/($km^{-3}\cdot h^{-1}$)	0.1649	0.2080	0.2913	0.3140
1热轧BFG使用流量/($km^{-3}\cdot h^{-1}$)	0.3477	0.5381	0.7577	0.7900
1800冷轧BFG使用流量/($km^{-3}\cdot h^{-1}$)	0.9401	1.1609	1.2028	1.2855
4#高炉用LDG流量/($km^{-3}\cdot h^{-1}$)	0.2742	0.4652	0.5561	0.6159
1热轧LDG使用流量/($km^{-3}\cdot h^{-1}$)	0.4697	0.6779	0.7241	0.8138
1.2#焦炉COG流量/($km^{-3}\cdot h^{-1}$)	0.1835	0.2142	0.2693	0.2574
1热轧COG流量/($km^{-3}\cdot h^{-1}$)	0.8551	1.0609	1.1017	1.2035
1800冷轧COG使用流量/($km^{-3}\cdot h^{-1}$)	0.8518	1.1852	1.2100	1.2703

情况 2 数据连续缺失.

在连续大量数据缺失的情况下, 本文同样选用以上的几种方法与本文方法进行对比. 本文分别考虑2013年8月30日的数据有3小时、6小时、10小时的连续缺失情况, 实验效果如图6和表3所示.

从表3可以看出, 对于生产情况比较稳定的用户, 例如高炉和焦炉, 采用本文方法和欧氏距离方法都会好于时间序列方法, 但是在用户的生产情况变化较大时, 由于欧式距离方法仅仅从相似的角度寻找相关数据, 从而使最终的模型训练不够准确, 数据填补效果变差. 相比之下, 基于MVWIC的相关数据分析可以更好的判断相似工况下的生产数据, 从而使数据填补具有更高的精度. 同时, 随着数据缺失的长度增加, 由于误差累积等原因, 几种方法的精度均有所下降, 但是本文方法的误差下降速度相对其他方法更为平缓.

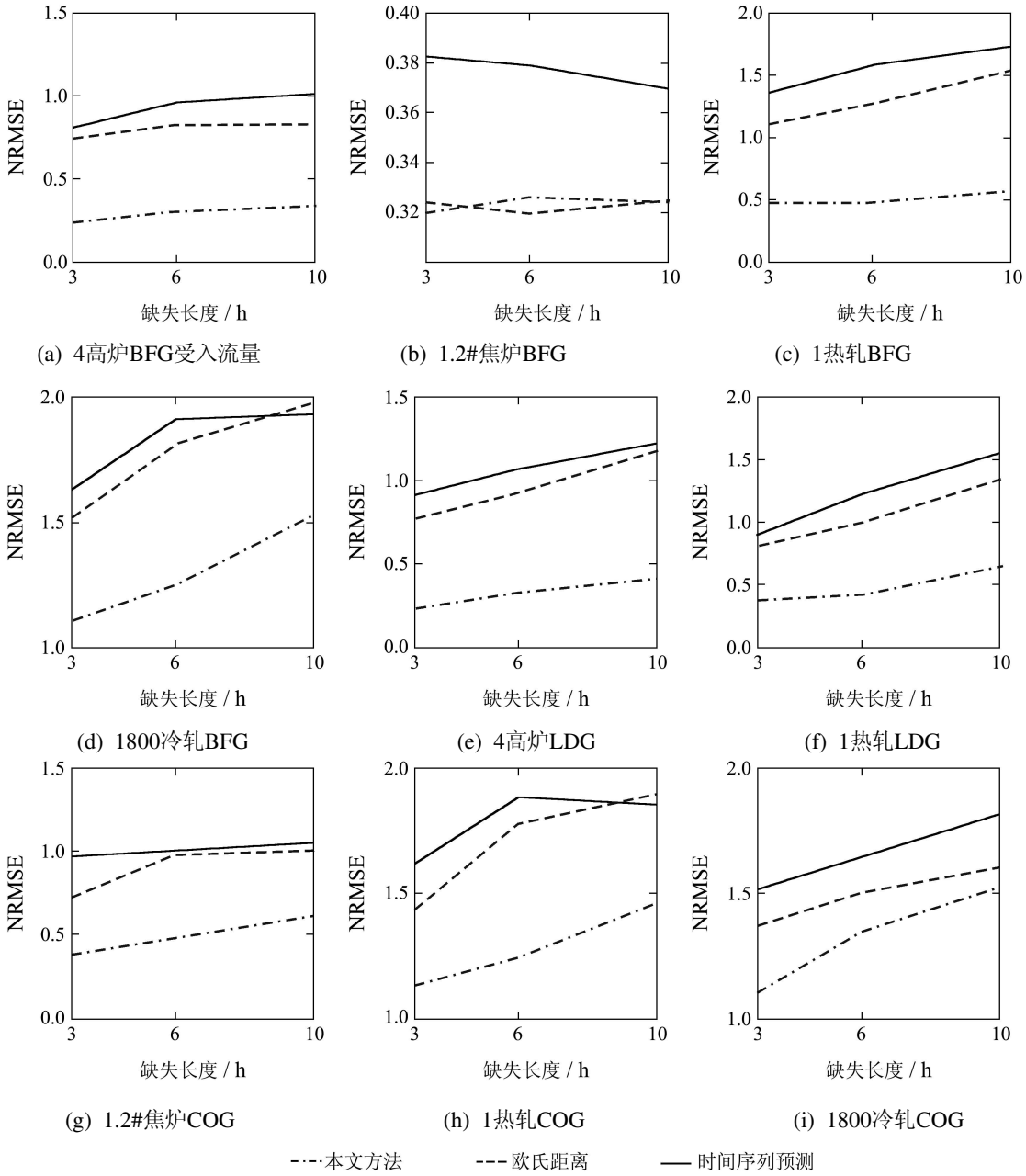


图 6 连续缺失数据的填补精度

Fig. 6 Imputation accuracy for consecutive missing data

表 3 不同缺失长度下的平均填补精度NRMSE

Table 3 Average NRMSE of different missing length

用户名称	本文方法	时间序列预测	欧式距离
4#高炉BFG发生流量/($km^{-3}\cdot h^{-1}$)	0.2902	0.9271	0.7978
1.2#焦炉BFG流量/($km^{-3}\cdot h^{-1}$)	0.3230	0.3771	0.3223
1热轧BFG使用流量/($km^{-3}\cdot h^{-1}$)	0.5040	1.5562	1.2994
1800冷轧BFG使用流量/($km^{-3}\cdot h^{-1}$)	1.2998	1.8253	1.7712
4#高炉用LDG流量/($km^{-3}\cdot h^{-1}$)	0.3263	1.0678	0.9592
1热轧LDG使用流量/($km^{-3}\cdot h^{-1}$)	0.4804	1.2246	1.0565
1.2#焦炉COG流量/($km^{-3}\cdot h^{-1}$)	0.4902	1.0072	0.9016
1热轧COG流量/($km^{-3}\cdot h^{-1}$)	1.2767	1.7833	1.6993
1800冷轧COG使用流量/($km^{-3}\cdot h^{-1}$)	1.3217	1.6556	1.4908

5 结语(Conclusions)

本文主要针对冶金工业生产过程中的煤气流量缺失值填补问题进行了讨论和实验. 将基于最大方差权信息系数的回归建模方法引入缺失值填补领域. 最大方差权信息系数将两个变量的分布用网格划分, 并用两组数据的互信息进行评定, 因此可以更准确的描述变量之间的关系, 同时, 由于方差权重项的引入, 对于冶金煤气流量数据这种噪声较大并且判定长度较小的情况, 可以更准确的找到具有相似工况的数据段. 使用同种工况下的数据进行回归模型更加符合现场的实际情况, 大大减小了填补误差. 实验结果证明, 本文的方法无论针对离散的缺失数据还是连续的数据缺失都有很好的效果, 为后期煤气管网平衡调度方案的制定以及实绩分析、编写生产报表等工作提供了有力的支持.

参考文献(References):

[1] LIU Y, LIU Q, WANG W, et al. Data-driven based model for flow prediction of steam system in steel industry [J]. *Information Sciences*, 2012, 193: 104 – 114.

[2] ZHAO J, WANG W, LIU Y, et al. A two-stage online prediction method for a blast furnace gas system and its application [J]. *IEEE Transactions on Control Systems Technology*, 2011, 19(3): 507 – 520.

[3] ZHAO J, LIU Y, ZHANG X, et al. A MKL based on-line prediction for gasholder level in steel industry [J]. *Control Engineering Practice*, 2012, 20(6): 629 – 641.

[4] LIU Y, ZHAO J, LV Z, et al. Study on balance scheduling based on the gasholder level prediction for blast furnace gas system in steel industry [J]. *Applied Mechanics and Materials*, 2012, 128: 1464 – 1467.

[5] LIU Y, ZHAO J, WANG W. A time series based prediction method for a coke oven gas system in steel industry [J]. *ICIC Express Letters*, 2010, 4(4): 1373 – 1378.

[6] ZHANG X, ZHAO J, WANG W. A improved multiple kernel learning based least square support vector regression and its application in the on-line prediction of gasholder level in iron and steel [J]. *ICIC Express Letter*, 2010, 4(5(B)): 1767 – 1772.

[7] DONALD B R. Inference and missing data [J]. *Biometrika*, 1976, 63(3): 581 – 592.

[8] RODERICK J A. Regression with missing x's: a review [J]. *Journal of the American Statistical Association*, 1992, 87(420): 1227 – 1237.

[9] KURODA M, SAKAKIHARA M. Accelerating the convergence of the EM algorithm using the vector algorithm [J]. *Comutational Statistics and Data Analysis*, 2006, 51(3): 1549 – 1561.

[10] GELMAN A, HILL J. *Data Analysis Using Regression and Multilevel/Hierarchical Models* [M]. London: Cambridge University Press, 2007.

[11] ANDRIDGE R R, RODERICK J A. A review of hot deck imputation for survey non-response [J]. *International Statistical Review*, 2010, 78(1): 40 – 64.

[12] SCHAFER J L, MAREN K O. Multiple imputation for multivariate missing-data problems: a data analyst's perspective [J]. *Multivariate Behavioral Research*, 1998, 33(4): 545 – 571.

[13] BOOTH J G, HOBERT J P. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm [J]. *Journal of the Royal Statistical Society*, 1999, 61(1): 265 – 285.

[14] SENTAS P, ANGELIS L. Categorical missing data imputation for software cost estimation by multinomial logistic regression [J]. *The Journal of Systems and Software*, 2006, 79(3): 404 – 414.

[15] GARCíA-LAENCINA P J, SANCHO-GóMEZ L, FIGUEIRAS-VIDAL A R, et al. K nearest neighbours with mutual information for simultaneous classification and missing data imputation [J]. *Neurocomputing*, 2009, 72(7): 1483 – 1493.

[16] TEMPL H M, FILZMOSER P. Imputation of missing values for compositional data using classical and robust methods [J]. *Computational Statistics and Data Analysis*, 2010, 54(12): 3095 – 3107.

[17] WANG Y, WAN W, WANG R, et al. Model, properties and imputation method of missing SNP genotype data utilizing mutual information [J]. *Journal of Computational and Applied Mathematics*, 2009, 229(1): 168 – 174.

[18] CHOONG M K, CHARBIT M, YAN H. Autoregressive-model-based missing value estimation for DNA microarray time series data [J]. *IEEE Transaction on Information Technology in Biomedicine*, 2009, 13(1): 131 – 137.

[19] BUUREN S V. Multiple imputation of discrete and continuous data by fully conditional specification [J]. *Statistical Methods in Medical Research*, 2007, 16(3): 219 – 242.

[20] LITTLE R J A, RUBIN D B. *Statistical Analysis with Missing Data* [M]. Second edition. New York: John Wiley & Sons, 2002.

[21] WAYNE F V, SUZANNE M C. A comparison of missing-data procedures for Arima time-series analysis [J]. *Educational and Psychological Measurement*, 2005, 64(5): 596 – 615.

[22] SUN H, PENG Y, CHEN J. A new similarity measure based on adjusted Euclidean distance for memory-based collaborative filtering [J]. *Journal of Software*, 2011, 6(6): 993 – 1000.

[23] NAGAI T, HOSHINO T, UCHIKAWA K. Statistical significance testing with mahalanobis distance for thresholds estimated from constant stimuli method [J]. *Seeing and Perceiving*, 2011, 24(2): 91 – 124.

[24] SON Y S, BAEK J. A modified correlation coefficient based similarity measure for clustering time-course gene expression data [J]. *Pattern Recognition Letters*, 2008, 29(3): 232 – 242.

[25] RESHEF D N, RESHEF Y A, FINUCANE H K, et al. Detecting novel associations in large data sets [J]. *Science*, 2011, 334(6062): 1518 – 1524.

[26] JORDAN M, KLEINBERG J, SCHOLKOPF B. *Pattern Recognition and Machine Learning* [M]. New York: Springer, 2006.

[27] LIITIAINEN E, VERLEYSEN M, CORONA F, et al. Residual variance estimation in machine learning [J]. *Neurocomputing*, 2009, 72(16): 3692 – 3703.

作者简介:

吕 政 (1987–), 男, 博士研究生, 主要研究方向为流程工业建模与优化调度, E-mail: lvzheng@mail.dlut.edu.cn;

赵 珺 (1981–), 男, 博士, 副教授, 主要研究方向为生产计划与调度、现代集成制造系统、工业生产一体化优化技术, E-mail: zhaoj@dlut.edu.cn;

刘 颖 (1982–), 女, 博士, 讲师, 主要研究方向为一体化生产计划与调度、系统仿真建模智能优化算法及应用、人工神经网络, E-mail: liu_ying@dlut.edu.cn;

王 伟 (1955–), 男, 教授, 博士生导师, 主要研究方向为自适应控制、现代集成制造系统和流程工业过程控制, E-mail: wangwei@dlut.edu.cn.