

基于不完整输入动态贝叶斯网的高炉消耗煤气量区间预测

陈龙¹, 刘颖¹, 赵琚¹, 王伟¹

1. 大连理工大学 控制科学与工程学院, 大连 116023

E-mail: liu_ying@dlut.edu.cn

摘 要: 针对钢铁企业煤气流量区间预测的问题, 考虑其数据噪声高和常含缺失值的特点, 本文提出了一种基于动态贝叶斯网建模的区间预测方法. 为了有效描述网络中节点间的非线性关系以及避免选择非线性基函数的个数和类型, 由核函数的线性叠加来构造节点间关系. 由于节点间的非线性关系, 在预测区间构造阶段, 采用了贝叶斯网近似推理的算法来进行区间推理. 为了验证所提方法的有效性, 测试了国内某钢铁厂高炉消耗煤气量的数据, 通过使几组模型输入按不同的比例随机缺失来构造不完整输入进行实验, 得出预测区间, 其结果表明了此动态贝叶斯网处理不完全输入情况的有效性.

关键词: 动态贝叶斯网, 区间预测, 核函数, 不完整输入

Prediction intervals for blast furnace consumption gas with incomplete input based on dynamic Bayesian networks

Long Chen¹, Ying Liu¹, Jun Zhao¹ Wei Wang¹

1. Dalian University of Technology, School of Control Sciences and Engineering, Dalian 116023

E-mail: liu_ying@dlut.edu.cn

Abstract: Considering the problem of prediction intervals for gas flow in steel plant, and given the industrial data feature of high level noises and incomplete input, a high order dynamic Bayesian network (DBN)-based PIs construction method for industrial time series is proposed in this study. For avoiding to designating the amount and type of the basis functions in advance, a linear combination of kernel functions is designed to describe the relationships between the nodes in the network, and because of the non-linear relationship between the nodes, the approximate inference is employed in the reasoning stage to construct the intervals. To verify the performance of the proposed method, a real-world industrial data (the consumption flow of blast furnace gas) are employed. The results indicate the effectiveness of our proposed method for the PIs construction of the industrial data with incomplete input.

Key Words: Prediction intervals, Dynamic Bayesian network, Kernel, Incomplete input

1 引言

在钢铁工业高炉炼铁过程中会产生大量的煤气, 同时有众多的生产环节 (如轧钢、热风炉和电厂等) 会实时消耗高炉煤气. 煤气的产生和消耗的量又会实时发生变化, 从而导致煤气管网压力的不平衡, 因此对煤气产消量的实时预测对于指导煤气的平衡调度有重要意义. 但由工业现场采集的煤气流量数据通常都含有高噪声, 并且常有缺失值的情况, 这样就增加了准确预测的难度. 另一方面, 针对工业数据含有高噪声的情况, 现场操作人员除了需获得变量的实时预测值之外, 对预测精度的可靠性也越来越重视.

针对钢铁企业的副产煤气流量预测问题, 目前已有不少学者进行了研究. 最常用的预测方法是使用神经网络^[1]和最小二乘支持向量机^[2]等. 文[3]研究了钢

铁企业能源系统中的副产煤气流量预测, 提出一种快速的参数优化的最小二乘支持向量机算法以适应在线预测的要求. 文[4]提出了基于数据驱动的贝叶斯回声状态网络方法来对蒸汽流量时间序列进行了预测. 文[5]提出结合影响高炉煤气发生流量的因素, 使用一种基于改进BP网络来构建高炉煤气发生量预测模型. 以上方法都是基于点的预测的模式, 这类方法只能给出预测的结果, 而不能说明结果的可靠性, 这增加了现场调度人员调度操作的风险. 而区间预测不仅能够给出预测结果, 同时能够给出与预测结果相应的可信度^[6]. 目前多采用Bootstrap方法、贝叶斯方法和delta方法等与神经网络相结合来进行预测区间的构造^[6]. 文[7]提出一种Bootstrap回声状态网络集成模型对高炉煤气系统的流量时间序列进行区间预测. 此外, 文[8]还提出一种基于核的在线区间预测构造方法. 然而以上点预测及区间预测方法都只是考虑了完整输入条件下的预测问题, 工业现场采集到的数据常含缺失值, 从而构成了不完整的输入, 使直接进行预测困难.

本文得到以下基金项目支持: 国家863计划(2013AA040703); 国家自然科学基金(61304213, 61273037, 61473056); 中央高校基本科研业务费专项资金(DUT13RC203).

动态贝叶斯网是贝叶斯网在时间上的扩展,它不仅表示了节点变量在同一时刻上的相互依赖关系,同时还表示了不同时刻节点的关系,能够更好地描述时序模型,并且容易处理缺失值^[9]. 对于连续变量贝叶斯网,目前已有线性高斯模型^[10,11]和非线性高斯模型^[12,13]等. 文[14]介绍了通过极大似然估计学习条件线性高斯模型的方法,但是煤气流量时间序列通常是复杂的非线性序列^[15]. 文[12]将动态贝叶斯网与非参数化回归结合起来描述基因网络的非线性关系,其中非线性基函数采用B样条函数,但需要指定B样条基函数个数,且需调整基函数的参数,过程繁琐.

本文考虑动态贝叶斯网框架能够方便地处理数据缺失问题,提出了一种基于动态贝叶斯网的建模方法,研究煤气流量数据存在缺失情况下时间序列的区间预测问题. 为了表示各节点间的非线性关系以及避免选择非线性基函数的个数和类型,本文提出采用核函数来表示贝叶斯网络中的节点间关系的方法. 考虑到节点间的非线性关系,该网络推理过程采用近似推理的方式来构造最后的预测区间. 为了验证本文方法应用在输入含缺失值的副产煤气流量区间预测问题的有效性,测试了国内某钢铁厂的高炉热风炉消耗煤气流量数据,通过使几组模型输入按不同的比例随机缺失来构造不完整输入进行实验,得出区间预测结果,验证了本文方法的有效性.

2 问题描述

钢铁企业副产煤气系统关系复杂,高炉、转炉、焦炉在生产过程中都会产生煤气,这些煤气进入各自的煤气管网,同时又有多个生产环节来消耗这些煤气. 图1展示了钢铁企业煤气系统. 从图1(a)可以看出,煤气系统分为发生、缓冲和消耗等环节. 发生环节包括高炉、转炉和焦炉;缓冲环节包括各管网的柜位及煤气管道;消耗环节众多,包括轧钢、化产、锅炉和电厂等. 图1(b)展示了高炉煤气发生的示意图,高炉产生煤气后其中一部分流量会先被高炉热风炉消耗而后进入管网,而每座高炉配有几座热风炉,经常存在换炉的情况,其消耗煤气量经常存在很大的波动,严重影响了煤气管网的平衡. 对高炉热风炉消耗煤气量进行实时预测对煤气管网的平衡调度有重要意义.

此外,在复杂的钢铁工业生产环境下可能会导致某些时刻采集的数据缺失,从而导致在线预测模型的输入不完整. 这样现有的区间预测方法无法直接使用不完整输入进行预测,需事先进行数据填补. 动态贝叶斯网由于同时表示了所有变量的依赖关系,在输入数据缺失时无需数据填补也容易推理出预测结果.

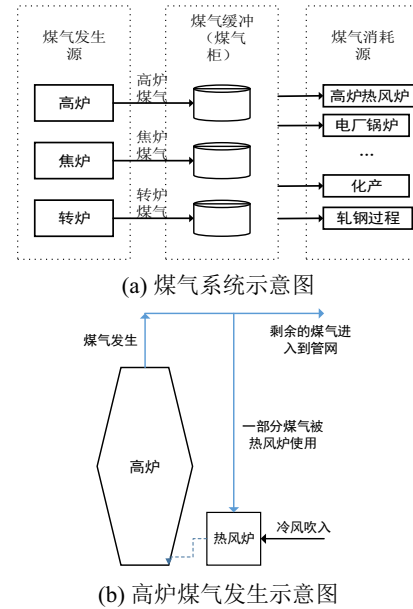


图1 副产煤气系统及高炉煤气发生示意图

3 基于核函数的动态贝叶斯网区间预测

工业数据普遍含有高噪声,而贝叶斯网络是一种概率图模型,与非线性回归结合更能够很好描述变量间的非线性关系. 文[12]研究了一阶非线性变量的动态贝叶斯网,然而一阶的模型不能很好的表达煤气流量时间序列的非线性关系,其区间预测需要更高阶的输入. 本文研究高阶的动态贝叶斯网络来进行工业数据的区间预测.

3.1 基于核函数的动态贝叶斯网

有工业时间序列数据集 $\{\mathbf{x}_i\}_{i=1}^N$, 其中 $\mathbf{x}_i = \{\mathbf{x}_{ij}\}_{j=1}^p = \{\mathbf{x}_i(t)\}_{t=1}^{l_i}$, $\mathbf{x}_{ij} = (x_{ij}(1), x_{ij}(2), \dots, x_{ij}(l_i))^T$, 并且 $\mathbf{x}_i(t)$ 表示第 i 条序列样本中第 t 个时间片上所有节点取值的向量, 每条序列样本长度是 l_i . 第 i 条序列可以表示为如下矩阵形式:

$$\begin{bmatrix} x_{i1}(1) & x_{i1}(2) & x_{i1}(3) & \dots & x_{i1}(l_i) \\ x_{i2}(1) & x_{i2}(2) & x_{i2}(3) & \dots & x_{i2}(l_i) \\ x_{i3}(1) & x_{i3}(2) & x_{i3}(3) & \dots & x_{i3}(l_i) \\ \dots & \dots & \dots & \dots & \dots \\ x_{ip}(1) & x_{ip}(2) & x_{ip}(3) & \dots & x_{ip}(l_i) \end{bmatrix}$$

同样, 数据集中所有第 t 个时间片上第 j 个节点值可以表示为 $\mathbf{x}_{[j]}(t)$. 假设随机变量 $X_j(t)$ 表示第 t 个时间片上的第 j 个变量, 本文考虑 n 阶动态贝叶斯网络, 设置其父节点不仅在同一时间片上而且也在之前的时间片上. 这样其父节点的值集合表示为 $\{\mathbf{pa}_{ij}(t)\}_{i=1}^N$, 其中 $\mathbf{pa}_{ij}(t) = \{\mathbf{pa}_{ij}^{(t)}(t), \dots, \mathbf{pa}_{ij}^{(t-k)}(t), \dots, \mathbf{pa}_{ij}^{(t-n)}(t)\}$, 且 $\mathbf{pa}_{ij}^{(t-k)}(t) = \{\mathbf{pa}_{ij,m}^{(t-k)}(t)\}_{m=1}^{q_{j,t-k}}$. $\mathbf{pa}_{ij}^{(t-k)}(t)$ 是 $X_j(t)$ 在第

$t-k$ 个时间片上的父节点的值向量. $\mathbf{pa}_{[j]}(t)$ 表示数据集中第 t 个时间片上第 j 个节点的值集合.

假定网络的结构和参数都是时不变的, n 阶动态贝叶斯网表示为如下概率密度的形式:

$$f_{DBN}(\mathbf{x}_i | \boldsymbol{\theta}_G) = \prod_{i=1}^{l_i} \prod_{j=1}^p f_j^{<\cdot>}(x_{ij}(t) | \mathbf{pa}_{ij}(t); \theta_j(t)), \quad (1)$$

$$i = 1, 2, \dots, N$$

其中, $\boldsymbol{\theta}_G = \{\boldsymbol{\theta}_G(t)\}_{t=1}^{n+1} = \{\boldsymbol{\theta}_{(i)}\}_{i=1}^p$, $\boldsymbol{\theta}_G(t) = \{\theta_i(t)\}_{i=1}^p$, $t = 1, 2, \dots, n+1$ 是 t 个时间片上参数的集合, 并且 $\theta_i(t)$ 表示了第 t 个时间片上第 i 个节点的参数. 条件概率密度函数 $f_j^{<\cdot>}(x_{ij}(t) | \mathbf{pa}_{ij}(t); \theta_j(t))$ 表示第 t 个时间片上第 j 个连续节点的概率密度. 因此, $f_j^{<\cdot>}(x_{ij}(t) | \mathbf{pa}_{ij}(t); \theta_j(t))$ 可以用高斯概率密度来表示, 如式(2)所示:

$$f_j^{<\cdot>}(x_{ij}(t) | \mathbf{pa}_{ij}(t); \theta_j(t)) = \frac{1}{\sqrt{2\pi\sigma_j^2(t)}} \exp\left[-\frac{(x_{ij}(t) - g_j^{<\cdot>}(\mathbf{pa}_{ij}(t)))^2}{2\sigma_j^2(t)}\right] \quad (2)$$

其中 $g_j^{<\cdot>}(\mathbf{pa}_{ij}(t))$ 表示节点 $X_j(t)$ 与其父节点之间的非线性关系. 这一非线性关系在文[12]中表示成基函数的线性组合, 见式(3).

$$g_j^{<\cdot>}(\mathbf{pa}_{ij}(t)) = \sum_{i=1}^M \omega_{ij}(t) \phi_{ij}^{<\cdot>}(\mathbf{pa}_{ij}(t)) = \boldsymbol{\omega}_j^T(t) \boldsymbol{\phi}_j^{<\cdot>}(\mathbf{pa}_{ij}(t)) \quad (3)$$

其中 $\boldsymbol{\omega}_j(t) = (\omega_{1j}(t), \omega_{2j}(t), \dots, \omega_{Mj}(t))^T$ 是权值向量, 并且 $\boldsymbol{\phi}_j^{<\cdot>}(\mathbf{pa}_{ij}(t)) = (\phi_{1j}^{<\cdot>}(\mathbf{pa}_{ij}(t)), \dots, \phi_{Mj}^{<\cdot>}(\mathbf{pa}_{ij}(t)))^T$ 是基函数的向量. 因此,

$$x_{ij}(t) = g_j^{<\cdot>}(\mathbf{pa}_{ij}(t)) + \varepsilon_{ij}(t), \quad (4)$$

$$i = 1, 2, \dots, N, j = 1, 2, \dots, p$$

其中噪声 $\varepsilon_{ij}(t)$ 服从均值为0的正态分布.

考虑到使用基函数线性回归的局限性, 使用基函数的数量需要根据输入空间的维数的增大而增大[16], 本文使用核函数而非基函数来表示节点间的非线性关系:

$$g_j^{<\cdot>}(\mathbf{pa}_{ij}(t)) = \sum_{k=0}^N \omega_{kj}(t) K_j^{<\cdot>}(\mathbf{pa}_{ij}(t), \mathbf{pa}_{kj}(t)) \quad (5)$$

其中, 当 $i=0$ 时 $K_j^{<\cdot>}(\mathbf{pa}_{ij}(t), \mathbf{pa}_{kj}(t)) = 1$.

基于以上的内容, 针对工业数据的基于核函数的动态贝叶斯网可以表示为式(6)形式:

$$f_{DBN}(\mathbf{x}_i | \boldsymbol{\theta}_G) = \prod_{i=1}^{l_i} \prod_{j=1}^p f_j^{<\cdot>}(x_{ij}(t) | \mathbf{pa}_{ij}(t), \theta_j(t^{n(t)})), i = 1, 2, \dots, N \quad (6)$$

其中, $n(t) = \begin{cases} 1, & t \leq n+1 \\ \log_t(n+1), & t > n+1 \end{cases}$. 由于所建网络是时不变的, 第 n 个时间片上的节点参数和结构之后时间片上的节点参数和结构相同.

3.2 基于核函数的动态贝叶斯网络参数学习

对于此动态贝叶斯网中的参数 $\boldsymbol{\omega}_j(t)$ 和 $\sigma_j^2(t)$, 本文采用极大似然估计来确定这些参数. 首先给出网络的对数似然函数, 如式(7)所示.

$$\ln f(\mathbf{x} | \mathbf{pa}(\mathbf{x}), \boldsymbol{\omega}, \boldsymbol{\sigma}) = \sum_{j=1}^p \sum_{t=1}^l \ln f(x_{[j]}(t) | \mathbf{pa}_{[j]}(t), \boldsymbol{\omega}_j(t), \sigma_j^2(t)) \quad (7)$$

由于网络的对数似然函数的是由网络中每个节点的似然度加和而成, 我们只考虑其中的子结构. 第 t 个时间片上第 j 个节点变量的似然函数表示为

$$\ln f(\mathbf{x}_{[j]}(t) | \mathbf{pa}_{[j]}(t), \boldsymbol{\omega}_j(t), \beta_j(t)) = \sum_{i=1}^N \ln N(x_{ij}(t) | \boldsymbol{\omega}_j(t)^T \mathbf{K}_j^{<\cdot>}(\mathbf{pa}_{ij}(t), \mathbf{pa}_{kj}(t)), \beta_j(t)) \quad (8)$$

$$= \frac{N}{2} \ln \beta_j(t) - \frac{N}{2} \ln(2\pi) - \beta_j(t) E_D(\boldsymbol{\omega}_j(t))$$

其中, $\mathbf{K}_j^{<\cdot>} = \mathbf{K}_j^{<\cdot>}(\mathbf{pa}_{ij}(t), \mathbf{pa}_{kj}(t))$, $\beta_j(t) = (\sigma_j^2(t))^{-1}$,

误差均方加和函数 $E_D(\boldsymbol{\omega}_j(t))$ 表示为

$$E_D(\boldsymbol{\omega}_j(t)) = \frac{1}{2} \sum_{i=1}^N \{x_{ij}(t) - \boldsymbol{\omega}_j(t)^T \mathbf{K}_j^{<\cdot>}\}^2 \quad (9)$$

为了求出使似然函数最大化的权值 $\boldsymbol{\omega}_j^{\text{ML}}(t)$, 本文求出似然函数对 $\boldsymbol{\omega}_j(t)$ 的偏导:

$$\frac{\partial (\ln f(\mathbf{x}_{[j]}(t) | \mathbf{pa}_{[j]}(t), \boldsymbol{\omega}_j(t), \beta_j(t)))}{\partial \boldsymbol{\omega}_j(t)} = \sum_{i=1}^N \{x_{ij}(t) - \boldsymbol{\omega}_j(t)^T \mathbf{K}_j^{<\cdot>}\} (\mathbf{K}_j^{<\cdot>})^T \quad (10)$$

将此偏导设置为0, 可以得到

$$0 = \sum_{i=0}^N x_{ij}(t) (\mathbf{K}_j^{<\cdot>})^T - \boldsymbol{\omega}_j(t)^T \left(\sum_{i=0}^N \mathbf{K}_j^{<\cdot>} (\mathbf{K}_j^{<\cdot>})^T \right) \quad (11)$$

求解式(11)可得 $\boldsymbol{\omega}_j(t)$ 的值为

$$\boldsymbol{\omega}_j^{\text{ML}}(t) = ((\boldsymbol{\Phi}_j^{<\cdot>})^T \boldsymbol{\Phi}_j^{<\cdot>})^{-1} (\boldsymbol{\Phi}_j^{<\cdot>})^T \mathbf{x}_{[j]}(t) \quad (12)$$

其中 $\boldsymbol{\Phi}_j^{<\cdot>}$ 是一个 $N \times (N+1)$ 的矩阵, 其第 k 行第 l 列的元素表示为 $\mathbf{K}_j^{<\cdot>}(\mathbf{pa}_{ij}(t), \mathbf{pa}_{kj}(t))$.

同样我们可以通过对 $\beta_j(t)$ 求偏导使之等于0, 可以求出 $\beta_j^{\text{ML}}(t)$ 使似然函数最大化, 可以得到

$$\frac{1}{\beta_j^{\text{ML}}(t)} = \frac{1}{N} \sum_{i=1}^N \{x_{ij}(t) - \boldsymbol{\omega}_j^{\text{ML}}(t)^T \mathbf{K}_j^{<\cdot>}\}^2 \quad (13)$$

因此, 最终可以求得 $\boldsymbol{\omega}_j^{\text{ML}}(t)$ 和 $\beta_j^{\text{ML}}(t)$ 是似然函数最大化.

3.3 预测区间构造及建模过程

对于工业数据预测问题,调度人员不仅重视预测结果本身还关心预测结果的可靠性,因此本文为预测结果构造预测区间.预测区间可以描述出预测结果带来的风险,假定 μ_{query} 表示预测变量的均值, σ_{query} 表示预测值的方差.由于网络中节点的非线性关系,本文首先将一个有 T 个时间片的动态贝叶斯网络在时间上展开成静态的网络,然后采用似然加权近似推理算法^[17]来推理进行未知变量节点的概率密度(μ_{query} 和 σ_{query}).则置信度为 $(1-\alpha\%)$ 的预测区间上下限分别由式(14)和(15)表示.

$$U_{limit} = \mu_{query} + z^{1-\frac{\alpha}{2}} \sigma_{query} \quad (14)$$

$$L_{limit} = \mu_{query} - z^{1-\frac{\alpha}{2}} \sigma_{query} \quad (15)$$

其中 U_{limit} 和 L_{limit} 分别为预测区间的上限和下限, $z^{1-(\alpha/2)}$ 是标准正态分布的 $1-(\alpha/2)$ 分位数.

由于煤气数据序列是一维时间序列,在动态贝叶斯网每个时间片上只对应于一个节点.这样此动态贝叶斯网的结构学习问题就转化为选择前 n 个节点为此节点的父节点.需要优化的参数还有核函数的参数,本文将核函数参数的优化和网络阶数 n 的选择同时进行.下面是区间预测模型建立的步骤:

步骤1: 分别设置动态贝叶斯网阶数的上下限为 n_{min} 和 n_{max} .

步骤2: 对于 n_{min} 和 n_{max} 之间的每一个阶数 n_{each} , 建立一个 n_{each} 阶动态贝叶斯网.其核函数的参数使用模拟退火优化算法和 L 折交叉验证来优化,优化目标是均方根预测的误差最小.

步骤3: 选取步骤2中得到这些贝叶斯网候选模型中交叉验证误差最小的模型作为最优模型.

步骤4: 对最优网络在时间上展开成静态网络,给定一组含缺失值的煤气流量输入,对此静态网络进行似然加权近似推理,通过式(14)和(15)进一步得到煤气流量预测区间的上下限.

4 数据试验及分析

为验证所提出的基于核函数的动态贝叶斯网区间预测方法的有效性,选取宝钢能源中心的煤气系统2015年10月份的高炉热风炉消耗高炉煤气流量数据进行验证,数据采样周期为1分钟.在实验中,分别对几组不同缺失比例的输入测试其区间预测的效果.

为了量化区间预测的效果,引入以下两个量化指标:根均方误差(RMSE)和基于宽度的覆盖指标(CWC).根均方误差表示为:

$$RMSE = \sqrt{\sum_{i=1}^n (Y_i - F_i)^2 / n} \quad (16)$$

其中, n 是预测的数据点数, Y_i 是观测值, F_i 是预测均值. RMSE用来表示预测均值的误差情况.基于宽度的覆盖指标定义如下^[6]:

$$CWC = NMPIW(1 + \gamma(PICP) \exp(-\eta(PICP - \mu))),$$

$$\gamma = \begin{cases} 0, & PICP \geq \mu \\ 1, & PICP < \mu \end{cases} \quad (17)$$

$$PICP = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} c_i \quad (18)$$

$$NMPIW = \frac{1}{R} \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (U_i - L_i) \quad (19)$$

其中,当目标值在预测范围内时 c_i 等于1,反之, c_i 等于0;而 U_i 和 L_i 分别是预测每个点的上限和下限, R 表示预测值的范围. $PICP$ 是预测区间的覆盖概率, $NMPIW$ 是归一化的平均区间宽度. η 和 μ 是控制 CWC 跳变点以及跳变幅度的两个超参数,本文取 $\mu = 10$, $\eta = 0.95$. CWC 指标值越小,表明区间预测的效果越好.

高炉热风炉消耗煤气流量序列样本集共含500条序列,每条序列长度为80.核函数采用高斯核函数 $K(\mathbf{x}, \mathbf{x}_i) = \exp(\|\mathbf{x} - \mathbf{x}_i\|^2 / 2b^2)$,在模型建立过程中采用了10折交叉验证以及使用模拟退火算法来优化核函数的参数 b^2 .最终得到在阶数为50时,交叉验证误差最小,此时得到的最优的核函数参数 b^2 为565.34.构建一个50阶动态贝叶斯网络,如图2所示,其中,第51时刻以后的节点的参数共享第51时刻节点的结构和参数.

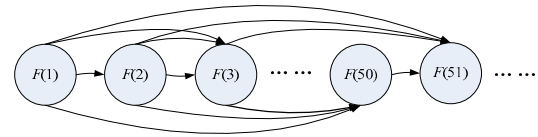
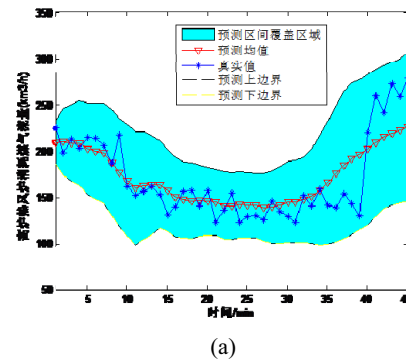
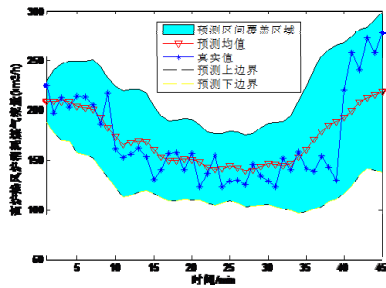


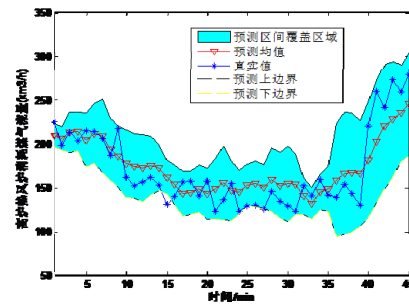
图2: 高炉热风炉消耗煤气流量预测的50阶动态贝叶斯网



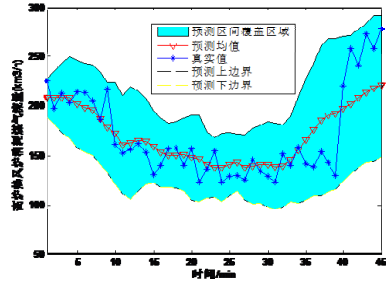
(a)



(b)



(d)



(c)

图3 高炉热风炉消耗煤气流量在不同比例输入缺失条件下的预测效果. (a)输入完整 (b)缺失5% (c)缺失10% (d)缺失30%

随机选取输入完整、输入随机缺失5%、10%和30%这四种输入情况的测试序列样本, 使用似然加权推理算法对图3中贝叶斯网进行推理得出相应的预测结果, 其中在似然加权算法中将样本量设为500. 将1至50时刻的节点值作为模型的输入来预测51至95

表 1 各方法在四种不同输入缺失比例情况下对高炉热风炉消耗煤气流量数据的预测结果对比

输入缺失比例	预测方法	CWC	PICP	NMPIW	RMSE (Km^3/h)		
输入完整	MLP	15.7709	0.6711	0.1871	21.1679		
	ESN	2.0783	0.8144	0.3595	26.1676		
	本文方法	1.5936	0.8778	0.4882	33.4647		
缺失5%	MLP	临近点均值填补	20.4503	0.6644	0.1867	21.8826	
		均值填补	4.8712	0.7044	0.2182	23.6847	
		样条插值	19.0812	0.6844	0.1861	20.9849	
	ESN	临近点均值填补	2.1444	0.8088	0.3589	26.3169	
		均值填补	2.5151	0.8055	0.3579	25.7350	
		样条插值	2.1913	0.8077	0.3589	26.2082	
	本文方法		1.6930	0.8700	0.4872	33.4788	
	缺失10%	MLP	临近点均值填补	19.7924	0.6733	0.1860	20.9724
			均值填补	14.0332	0.7177	0.2594	29.9994
样条插值			19.8487	0.6644	0.1841	21.8767	
ESN		临近点均值填补	2.2483	0.8077	0.3589	26.2693	
		均值填补	2.0299	0.8188	0.3584	25.3482	
		样条插值	2.2396	0.8077	0.3591	26.3640	
本文方法		1.7150	0.8656	0.4721	33.4051		
缺失30%	MLP	临近点均值填补	20.6290	0.6488	0.1839	22.5021	
		均值填补	12.9045	0.7000	0.3250	38.0552	
		样条插值	20.9042	0.6511	0.1824	21.4786	
	ESN	临近点均值填补	2.5339	0.7988	0.3567	26.5444	
		均值填补	3.1198	0.8111	0.3525	26.4455	
		样条插值	2.2241	0.8077	0.3581	25.8463	
	本文方法		2.1838	0.8067	0.4337	33.8752	

时刻的节点的概率分布. 预测结果如图3所示, 图中给出了95%的置信区间. 从图3中可以看出随着输入缺失比例的增大, 其预测质量有所下降, 但真实数据值却很少超出其预测区间的范围.

为了进一步说明本文方法的有效性, 将本文方法的预测结果与基于贝叶斯方法的多层感知神经网络(MLP)和基于bootstrap方法的回声状态网络(ESN)的预测结果分别进行了对比. 将数据填补后的缺失输入

作为对比方法的输入,分别使用了邻近点均值、均值和样条插值三种数据填补方法.各方法的区间预测评价结果统计见表1(统计了30次随机实验的平均结果),从表1中可以看出本文方法具有最高的精度.从CWC指标来看,对于四种输入缺失情况,虽然对于RMSE本文方法不是最小的,但其CWC值却都是最小的,表明其区间预测的质量是最好的.表1中还列出了PICP和NMPIW两个指标的值.在输入完整、缺失5%和10%的情况下,本文方法的PICP都是最大的,表明其预测区间覆盖的概率是最大的.其他两种对比方法在同一输入缺失比例的情况下,不同的数据填补方法会导致其预测效果存在很大的差异,说明数据填补质量的不同对预测结果有很大影响.此外,图3(d)展示了在高输入缺失比例(缺失30%)下的预测结果,其中有部分真值超出了预测区间的范围,预测质量较低缺失比例时有所下降.在高缺失比例下,对于CWC、PICP和NMPIW等各项预测指标,本文方法虽不总是最优的,但其预测结果却是相对稳定的(对比方法中基于样条插值填补的ESN方法的预测结果表现最佳,但其他填补方法下的预测质量却参差不齐).综上所述,使用本文方法进行区间预测在处理缺失值输入时存在如下优势:无需事先选择合适的填补方法直接对缺失输入预测,其区间预测的稳定性更高.

5 结论

本文针对钢铁企业中高炉热风炉消耗煤气流量的区间预测问题,提出使用基于函数的动态贝叶斯网的方法来构造预测区间.节点间的非线性关系提出采用核函数的线性叠加来表示.本文采用似然加权推理算法进行概率推理并构造预测区间.实验表明,使用本文方法进行缺值输入的区间预测,会得到较稳定的区间预测效果.此方法对钢铁企业高炉煤气的调度问题有很好的指导意义.

致谢:

国家自然科学基金(No. 61273037, 61304213, 61473056, 61533005, 61522304, U1560102), 国家863计划(No. 2013AA040703), 国家科技支撑计划(No. 2015BAF22B01)和中央高校基本科研业务费(DUT15YQ113).

参考文献

- [1] Bhaskar K. and Singh S. N. A WNN-assisted Wind Power Forecasting Using Feed-forward Neural Network. IEEE Transactions on Sustainable Energy, Vol. 3, No.2: 306-315, 2012.
- [2] 阎威武, 邵惠鹤. 支持向量机和最小二乘支持向量机的比较及应用研究. 控制与决策. Vol. 18, No.3, 358-360, 2003.
- [3] Zhao J., Liu Q.L., Pedrycz W., et al. Effective Noise Estimation-Based Online Prediction for Byproduct Gas System in Steel Industry. IEEE Transactions on Industrial Informatics. Vol. 8, No. 4, 953-963, 2012.
- [4] Liu Y., Liu Q.L. and Wang W., et al. Data-driven based Model for Flow Prediction of Steam System in Steel Industry. Information Sciences. Vol. 193, 104-114, 2012.
- [5] 肖冬峰, 杨春节, 宋执环. 基于改进BP网络的高炉煤气发生量预测模型. 浙江大学学报:工学版. Vol. 11, 2103-2108, 2012.
- [6] Khosravi A., Nahavandi S. and Creighton D., et al. Comprehensive Review of Neural Network-based Prediction Intervals and New Advances. IEEE Transactions on Neural Networks. Vol. 22, No. 9, 1341-1356, 2011.
- [7] Sheng C., Zhao J., Wang W., et al. Prediction Intervals for A Noisy Nonlinear Time Series based on A Bootstrapping Reservoir Computing Network Ensemble., IEEE Transactions on Neural Networks and Learning Systems. Vol. 24, No. 7, 1036-1048, 2013.
- [8] 赵珺, 杜雅楠, 盛春阳,等. 基于核的冶金煤气流量在线区间预测. 控制理论与应用. Vol. 10, 1274-1280, 2013.
- [9] Murphy K. P. Dynamic Bayesian Networks: Representation, Inference and Learning. University of California, Berkeley, 2002.
- [10] Jordan M. I. An Introduction to Probabilistic Graphical Models. Preparation, Vol. 42, No.2, 27-56. 2003.
- [11] Dan G., Heckerman D. Learning Gaussian Networks. Uncertainty Proceedings. 235-243, 1994:
- [12] Kim S., Imoto S. and Miyano S. Dynamic Bayesian Network and Nonparametric Regression for Nonlinear Modeling of Gene Networks from Time Series Gene Expression Data. Biosystems Vol. 75, No.1, 57-65, 2004.
- [13] Imoto S., Kim S. and Goto T., et al. Bayesian Network and Nonparametric Heteroscedastic Regression for Nonlinear Modeling of Genetic Network. Journal of bioinformatics and computational biology, Vol.1, No.02, 231-252. 2003,
- [14] Murphy K. P. Fitting A Conditional Linear Gaussian Distribution. 1998.
- [15] Liu Y., Zhao J. and Wang W., et al. Study on Prediction Method for Generation and Consumption of Coke Oven Gas. 2010 8th World Congress on Intelligent Control and Automation, 4446-4451. 2010.
- [16] Bishop C. M. Pattern Recognition and Machine Learning. New York: Springer, 2006.
- [17] Fung R. and Chang K. C. Weighting and Integrating Evidence for Stochastic Simulation in Bayesian Networks. Uncertainty in Artificial Intelligence. Elsevier: Bonissone P. P., Henrion M., Kanal L. N., and Lemmer J. F, 208-219, 1990