文章编号: 1002-0411(2002)04-380-05

智能建模方法中的数据预处理

杨 斌 田永青 朱仲英

(上海交通大学自动化系 200030)

摘 要: 本文对智能系统建模中的数据预处理问题进行了研究, 指明了这一工作的重要性. 结合一个系统建模的例子阐明了在智能建模方法中进行辅助变量初选、数据采集、数据处理与校正、输入数据集降维等各种数据预处理工作的思路与方法, 为系统模型的准确建立提供了保证. *

关键词: 建模; 数据预处理; PCA; 神经网络; 燃烧系统

中图分类号: TP13 文献标识码: B

DATA PRETREATMENT IN INTELLIGENT MODELING METHODS

YANG Bin TIAN Yong-qing ZHU Zhong-ying

(*Department of Automation, Shanghai Jiaotong University, Shanghai* 200030, *China*)

Abstract: This paper researched on the data pretreatment problems in intelligent system modeling, and pointed out the importance of this work. With an example of system modeling, the paper expounded contents and methods in various phases of data pretreatment of intelligent modeling, such as the primary election of variable, data collection, data processing and correcting, reducing dimensions of the input data etc. The work gave security to the accurate setting-up of the system model.

Keywords: modeling, data pretreatment, PCA, neural network, burning system

1 引言(Introduction)

在工业控制和许多其它的应用领域, 建立系统的模型是非常重要的一个步骤. 目前常用的建模方法主要有三种: 机理建模方法、辨识建模方法、智能建模方法.

由于多数系统机理复杂, 难以完全从机理上揭示其内在规律, 系统的非线性、分布参数、时变性、时滞性等特点, 都给上述的前两种传统方法建模带来了困难. 近年来, 神经网络作为智能建模方法的代表, 以其强大的非线性拟合能力、并行信息处理能力和自学习能力, 而得到越来越多的应用. 神经网络建模所需要的信息全靠从训练样本中得到, 这就决定了建模效果的好坏依赖于样本的数量和质量[3], 因此对样本数据的预处理显得非常重要.

本文将以应用广泛的工业及民用供热(供暖)层燃型锅炉的燃烧系统为例子, 讨论对其进行神经网络建模过程中的数据预处理问题. 实践结果表明, 这些技术的应用取得了很好的建模效果.

2 神经网络建模的工程化设计、实施步骤(Design and execution steps of neural network modeling)

神经网络建模是实用性很强的应用技术, 因此其设计必须满足工程应用的简易性、有效性、可靠性要求. 神经网络模型的工程化设计、实施一般分以下几个步骤[1,2]:

(1) 辅助变量的初选

根据系统工作机理, 在可测的变量集中初步选择所有与被估计变量有关的原始辅助变量, 这些变量中部分可能是彼此相关的变量.

(2) 现场数据采集与处理

采集被估计变量和原始辅助变量的历史数据. 现场数据必须经过过失误差检测和数据协调, 保证数据的准确性. 由于神经网络建模一般用于静态估计, 应该采集系统平稳运行时的数据, 并注意纯滞后的影响.

(3) 辅助变量精选—— 输入数据集降维

通过机理分析, 可以在原始辅助变量中找出相关的变量, 选择响应灵敏、测量精度高的变量为最终

* 收稿日期: 2001-11-03

的辅助变量. 更为有效的方法是主元分析法, 即利用现场的数据作统计分析计算, 将原始辅助变量与被测量变量的关联度排序, 实现变量精选.

(4) 神经网络模型的结构选择

根据系统特点选择模型的类型, 即线性、非线性和混合型等.

(5) 模型参数的估计

利用样本数据对网络进行训练. 为了检验模型的有效性, 一般将历史数据集中分为两部分, 一部分用于参数估计, 另一部分用于模型检验. 若检验表明模型达到了预定精度, 即可将模型投用.

(6) 神经网络模型实施

利用软、硬件实现神经网络模型.

(7) 在线数据预处理

辅助变量的在线测量数据, 必须经过除噪滤波, 显著误差检测及数据校正, 方可作为神经网络模型的输入.

神经网络模型实施以后, 若计算结果与实际的离线测量值仍有误差, 则需进行模型校正.

以上各个步骤中, (4)、(5)、(6) 分别是神经网络模型的选型、训练与实施, 本文将不作讨论. (1)、(2)、(3)、(7) 涉及到神经网络模型建立过程中的许多数据处理问题, 将是下面讨论的重点.

3　辅助变量初选与数据采集(Assistant variables selection and data collection)

3.1　概述

层燃炉燃烧系统的排烟热损失控制一直是一个难点, 排烟氧含量是反映排烟热损失的一个重要参数. 对排烟热损失控制效果不理想的原因, 一方面有燃烧系统大滞后的因素, 还有一方面是对排烟氧含量进行检测存在困难. 目前测量氧含量的各种分析仪表操作复杂、价格昂贵、寿命短. 这些都对成功的控制增加了困难.

对于这一个复杂的非线性系统, 我们希望用神经网络建立一个模型, 对排烟氧含量进行实时的估计, 从而克服滞后和硬件测量困难, 为改进控制提供一条新途径.

3.2　辅助变量初选

我们要建立的是一个 M ISO 系统模型, 系统的输出是排烟氧含量, 系统的输入就是我们要选择的辅助变量. 辅助变量的初选通常是根据对系统的机理分析而进行的. 根据对锅炉燃烧系统机理的分析, 不难得到对排烟氧含量有影响的几个系统输入变量: 炉膛温度、炉排转速、鼓风机输出、引风机输出、炉膛负压. 我们的目标就是建立一个 5 输入 1 输出的神经网络模型.

3.3　数据采集

在对变量(包括 5 个辅助变量和 1 个输出变量)的数据进行采集时, 首先需要详细研究燃烧系统的各种操作工况, 确定各个变量可能的取值范围. 采集样本的空间要尽量覆盖整个操作范围, 且注意选择的每一个样本在样本空间内要有一定的代表性.

在整个样本空间内, 要选择合适的样本量, 样本数据要均匀. 切不可在样本空间的某一段选取大量重复的数据, 一方面不利于网络学习, 另一方面采集的样本数量大, 很难保证数据相互之间保持一致性. 不一致的数据可能是由于过程噪声或干扰影响, 也有可能属于过失误差, 将这些数据用于建模即使有好的训练精度也往往泛化结果不好. 类似的, 不可在样本空间的边缘只选取零星的几个数据, 这样采集的数据用于建模以后往往泛化结果很差.

当然, 若采集到的数据真正满足上述的均匀性、代表性、精简性的原则, 理论上讲样本越多, 越能更好的反应过程的特性.

本文是在某居民小区供热中心锅炉 2000 年 2 月份的历史数据记录中, 选取神经网络建模的训练样本数据. 由于控制系统的上位机中存储的大量数据因机器故障丢失, 因此由上位机整点打印的数据报表成为了采样的重要依据. 我们选择锅炉稳定工作、负荷较大的 2 月份进行数据采集, 每天按整点记录有 24 组数据, 共计 700 余组数据. 这些数据在上位机内部已经过一定的滤波处理, 滤去了一部分随机噪声的影响.

得到数据以后, 首先进行的工作是粗选, 这一过程主要通过手工识别和精简来完成. 通过粗选选出锅炉在各种工况下相对稳定运行时的测量数据, 也即在挑选数据时考虑了燃烧系统大滞后. 一般认为对于燃烧系统这样的大滞后、非线性复杂系统, 其动态过程数据是病态的, 这些数据对软测量模型的建立是有害的, 因此要注意剔除.

4　数据处理与校正(Data treatment and verification)

在对系统数据粗略处理的基础上, 进行进一步的数据处理与校正是一个不可缺少的环节. 剔除原始数据中的过失误差, 降低随机误差对采样值的影响, 是数据处理与校正技术要解决的问题.

在离线处理训练样本数据时, 关键是剔除样本数据中的过失误差. 我们根据工艺要求和操作经验, 总结出被采集变量的操作范围, 然后采用最大值最小值限幅的方法先初步剔除一部分不在此范围内的数据. 进一步我们根据研究的主要目标排烟氧含量进行排序, 然后结合锅炉燃烧机理知识、操作经验等, 分别考虑每一个辅助变量与输出变量的相互关系、作用方式等. 将一系列定性分析的结果作为规则, 将采集数据中与这些规则相违背的作为过失误差剔除. 比方说, 锅炉燃烧系统的炉膛温度对排烟氧含量影响很大, 炉膛温度越高则燃烧越充分, 通常排出烟气中的氧含量越低. 如果发现在一组样本中炉膛温度和排烟氧含量都相对较大, 那么说明这组样本可能存在过失误差或者这组样本是锅炉在工况变化时采集的病态数据, 应该予以剔除. 各输入输出变量的采样值均由上位机报表中获得, 因此人为产生过失误差的可能性基本没有. 产生的过失误差是由系统中的大干扰或测量仪表失灵而引发, 而这些数据往往偏差比较大, 容易通过限幅方法来剔除.

在处理完过失误差以后, 对于采集的样本数据内的随机误差, 一般通过加权取均值的滤波方法来处理.

在神经网络模型在线应用时, 对于采集的系统数据的过失误差和随机误差也必须进行处理. 在每一个模型预报周期内可以考虑三个方面: 判别系统平稳性、评定过失误差、处理随机误差. 根据现场模型预报所需的计算时间, 我们将锅炉燃烧系统排烟氧含量预报周期定为 30s. 下面我们以 7 点滑动平均方法为例.

在每一个预报周期内首先要判别系统的平稳性. 根据大量的历史数据和操作经验, 可以确定燃烧系统在稳态下各辅助变量的变化幅度. 在线运行时, 若辅助变量最近 7 个采样数据的变化幅度未超出范围, 可认定系统处于相对稳态. 下一步可以评定过失误差. 这里除了使用上面所述的根据工艺操作范围的最大最小值限幅的方法以外, 还可以根据操作情况, 定义一个辅助变量最大变化率, 若前后采集的两组数据的变化率超过此值, 应该作为过失误差剔除. 对于一个预报周期内的多组采样数据使用滑动平均法来处理, 即将一定时间内的系统测量值加权平均后使用, 可以消除大部分过程噪声的影响. 下面简单讨论一下随机误差的处理步骤.

尽管滑动平均法是一种古典的数据处理方法, 但由于其简捷性和较好的滤波作用, 在处理短时段内的随机误差时非常有效[3]. 设各辅助变量的动态测试数据可以描写为:

$$y_{\tilde{k}} = y_k + e_k \quad (k = 1, 2, L, N) \qquad (4\text{-}1)$$

其中 y_k 为采集信号中的确定性成分, $y_{\tilde{k}}$为实际测量值, e_k 为随机起伏的测量误差和噪声, k 为离散采样时刻, N 为区间内的数据长度. 为了抑制随机误差 $\{e_k\}$ 的影响, 就需对动态测量数据 $\{y_{\tilde{k}}\}$ 作平滑和滤波处理. 具体地说, 就是对非平稳的数据 $\{y_{\tilde{k}}\}$ 在适当的小区间上(设有 m 个相邻数据的小区间) 视为接近于平稳, 而作某种局部平均以减少随机误差 $\{e_k\}$ 所造成的随机起伏. 这样对全长 N 个数据逐一小区间上进行不断的局部平均, 即可得到较平滑的测量结果. 最简单的滑动平均方法就是将 m 个相邻数据直接作算术平均, 即等权平均处理的方法. 但由于在实际过程中一般是相距平滑均值 $\{y_{\tilde{k}}\}$ 较远的数据对平滑的作用可能要小于较近者, 因而一般应采用不等权的加权平均方法来滤波比较合理. 本文就采用了这一方法, 其一般算式为:

$$\{y_{\tilde{k}}\} = \sum_{i=-q}^{p} w_i y_{k+i}$$
$$(k = q + 1, q + 2, L, m - p) \qquad (4\text{-}2)$$

式中 w_i 为权系数, 且满足 $\sum w_i = 1$; p, q 为小于 m 的任意正整数, 且有 $p + q + 1 = m$. 一般平滑数据取在 m 个相邻数据的对称中点(取 $p = q$), 即中心平滑法. 这时其权系数 $\{w_i\}$ 是对称分布的, 若采用二项式分布的形式来分配其权系数, 即 $\{w_i\}$ 是二项式 $(1/2 + 1/2)^{2n}$ 展开的各项值, 式(4-2) 就变为:

$$\{y_{\tilde{k}}\} = \sum_{i=-n}^{n} w_i y_{k+i}$$
$$(k = n + 1, n + 2, L, m - n) \qquad (4\text{-}3)$$

其中有 $w_i = c_i/A$, $w_{-i} = w_i$, $A = 2^{2n}$, 这里采用 7 点中心平滑, 即 $m = 7$, $n = 3$, $A = 64$, $c_0 = 20$, $c_1 = 15$, $c_2 = 6$, $c_3 = 1$.

5 输入数据集降维—— PCA 算法(Debasement of input data dimensions- PCA arithmetic)

PCA (Principal Components Analysis) 即主元分析, 是一种统计相关分析技术[4]. 在实际生产中, 为了全面分析问题常常提出许多与输出有关的变量, 每个变量都在不同程度上反映了过程的某些信息, 但往往它们之间有一定的相关关系. 太多变量构成的高维数据空间使建模问题复杂化, 同时若众多的变量间还存在错综复杂的相关关系, 则又会给建

模带来困难. 所以基于这个特点, 在进行建模前, 通过 PCA 方法找出几个公共的支配因素, 最大限度保留有用信息, 滤去冗余信息, 然后按主元贡献率选取合适的主元数目进行过程建模, 将会大大简化模型结构和建模工作量[5,6].

考虑 X ($n\times m$ 矩阵) 为自由变量组成的矩阵, n 代表样本数组, m 代表变量个数. PCA 将矩阵 X 分解为如下形式:

$$X = t_1 p_1^{\mathrm{T}} + t_2 p_2^{\mathrm{T}} + \cdots + t_k p_k^{\mathrm{T}} + E \qquad (5\text{-}1)$$

(5-1) 式等价于:

$$X = TP^{\mathrm{T}} + E \qquad (5\text{-}2)$$

其中, t_i 为各主元系数; p_i^{T} 为矩阵 X 的协方差阵由大到小排列的第 i 个特征根所对应的特征向量, 称为主元向量, 包含各变量之间相互关联的信息; E 为主元素投影后与 X 的差值部分. 式中 k 必须小于或等于 X 的最小维数. 每一对 t_i, p_i 都是按相应于特征向量 p_i 的特征值 λ_i 的降幂排列, 其中第一对截获了所有分解的主元向量和主元系数对中最大的信息量, 其余以次类推. 并且满足:

$$Xp_i = t_i \quad \text{或者} \quad T = XP \qquad (5\text{-}3)$$

并且每一对都可以计算出其主元贡献率 η_i(如式 5-4 所示). 经验指出, 一般前 p 个主元的累计贡献率超过 85% 就认为 p 个主元反映了过程的主体信息.

$$\eta_i = \frac{\lambda_i}{\sum_{i=1}^{k} \lambda_i} \qquad (5\text{-}4)$$

PCA 的主要思想就是通过引入统计分析的方法, 降低数据冗余程度, 为后面的工作减少运算量, 节省硬件和时间. 具体的讲, PCA 算法的主要步骤可以概括为下面几步:

(1) 由自由变量按列组成原始数据矩阵 X.

(2) 求出 X 矩阵的协方差矩阵 S, 并求出 S 的特征值矩阵 V 和相应的特征向量矩阵 U.

(3) 求出矩阵 $T = X * U$.

(4) 分析特征值矩阵 V, 得到主元累计贡献率超过 85% 所需的的主元个数 p 的数值.

(5) 在 T 矩阵中选出相应于主元的 p 个向量, 作为神经网络训练的样本数据.

在本次燃烧系统建模中, 我们将主元分析方法与 RBF 神经网络相结合, 既保留了原始变量的特征信息, 又简化了神经网络结构, 使 RBF 网络的许多优点得以充分发挥[6,7]. 图 1 给出了原始数据经 PCA 处理后的主元变量作为 RBF 模型输入的结构示意图.

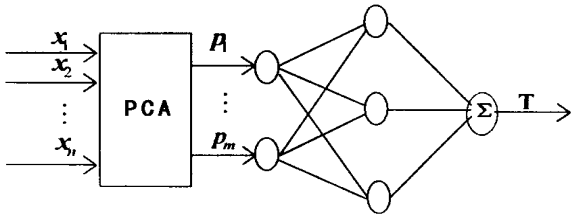

图 1　PCA-RBF 结构

Fig. 1　Structure of PCA-RBF

我们采集了锅炉燃烧系统的数据, 按照 PCA 的要求, 在对 RBF 网络进行训练以前, 首先对输入变量采样数据进行主元分析. 表 1 为主元分析的结果. 可以看出, 前两个主元的累计贡献率已经超过 85%, 说明只用两个主元就基本上可以反映五个输入变量所包含的系统输入信息. 根据 PCA 算法, 可以由辅助变量矩阵 X 和特征向量矩阵 P 得出主元变量 T, 下一步就可以用 T 作为网络输入对 RBF 网络进行训练. 可见, 有了主元分析, 将 5 输入 1 输出的 MISO 系统简化为了 2 输入 1 输出的 MISO 系统.

表 1　主元分析结果

Tab. 1　Results of PCA analysis

PC No.	Eigenvalue	Variance(%)	Total Variance(%)
1	0.0027	57.45	57.45
2	0.0016	34.04	91.49
3	0.0003	6.38	97.87
4	0.0001	2.13	100
5	0	0	100

6　仿真结果(Results of simulation)

仿真数据来源于某居民小区供热锅炉的实际运行数据, 这些数据反映了锅炉燃烧系统在各种工况下稳定运行时的的系统特性, 并且考虑了滞后的影响. 大量的实测数据在进行模型训练或泛化以前, 经过了过失误差检测、测量误差检验、滤波、归一化等处理, 选出具有均匀性、代表性、精简性的 55 组数据对模型进行训练, 35 组数据对模型进行泛化检验. 以下对模型的仿真使用了上述 90 组数据.

PCA-RBF 网络进行训练和泛化检验, 其结果见于图 2、3. 图中, 横坐标代表样本序号, 纵坐标是氧含量归一化值; 实线为实测输出数据, 虚线为网络训练或泛化输出数据. RBF 网络采用 OLS 算法, 对三层 RBF 网络进行训练, 训练耗时 3.1 s, 隐层神经元 54 个. 训练结果主要数据如下. 训练 RMSE: 0.

0012, 训练 MAXE: 0. 0056, 泛化 RMSE: 0. 0060, 泛化 MAXE: 0. 0114. (RMSE 指均方根误差; MAXE 指最大绝对误差)

$$RMSE = \sqrt{\frac{1}{z}\sum_{i=1}^{z}(y_i - y_{\hat{i}})^2} \qquad (6\text{-}1)$$

$$MAXE = \max_{i=1}^{z}(|y_i - y_{\hat{i}}|) \qquad (6\text{-}2)$$

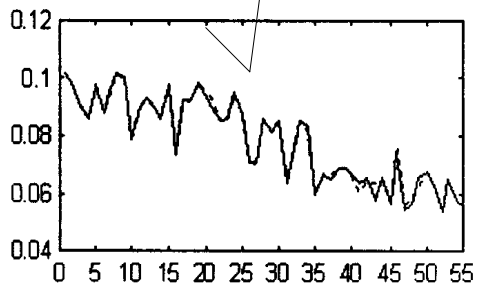

图 2 PCA-RBF 网络训练结果

Fig. 2 Training result of PCA-RBF network

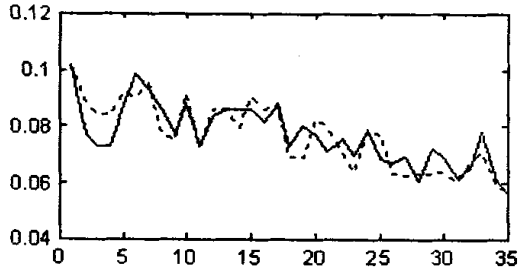

图 3 PCA-RBF 网络泛化结果

Fig. 3 Generalization result of PCA-RBF network

相对于本文所研究的锅炉对象而言, PCA 可以实现输入变量的降维, 但并不能够全面简化 RBF 网络的结构. 但是可以预见, 如果系统再复杂一些, 输入变量维数再高一些, 则 PCA-RBF 将可以大大简化网络结构, 从而较单纯的 RBF 具有明显优势.

7 结语(Conclusion)

本文对智能系统建模中的数据预处理问题进行了研究, 进一步指明了这一工作的重要性. 只有处理好包括辅助变量初选、数据采集、数据处理与校正、输入数据集降维等在内的各个环节, 才能为系统模型的简洁性、准确性提供保证. 仿真的结果有力的证明了这些思想和方法的有效性. 在这一领域中各种新方法层出不穷, 研究前景广阔.

参 考 文 献 (References)

1 孙 欣, 王金春, 何声亮. 过程软测量. 信息与控制, 1998, **27**(4)

2 于静江, 周春晖. 过程控制中的软测量技术. 控制理论与应用, 1996, **13**(2)

3 仲 蔚. 软测量与先进控制策略研究及其在石油化工中的应用. 华东理工大学博士学位论文, 1999, 12

4 JEDWARD JACKSON. Principal Components and Factor Analysis: Partl Principal Components. Journal of Quality Technology, 1980, **12**(4)

5 MICHAEL J, PIOVOSO and KARLENE A. KOSANOVICH. Applications of multivariate statistical methods to process monitoring and controller design. INT. J. CONTROL, 1994, **59**(3): 743~ 765

6 潘立登, 黄晓峰. 用 PCA-RBFN 建立可侦破故障的反应器自校正模型. 石油化工自动化, 1998, 1: 23~ 25

7 阳宪惠, 冯雄峰, 徐用龚. 油品质量估计中的统计分析方法. 化工自动化及仪表, 1996, **23**(6): 28~ 33

作者简介

杨 斌(1975-), 男, 博士研究生. 研究领域为地理信息系统、数据挖掘、智能控制等.

田永青(1974-), 男, 博士研究生. 研究领域为智能决策支持系统、地理信息系统、数据挖掘等. .

朱仲英, 男, 教授, 博士生导师. 研究领域为智能空间信息系统和智能空间决策支持系统的理论与应用研究、智能控制系统的理论与应用研究.