基于向量相似度的多模型局部建模方法研究*

曾 静[a], 王 军[b], 郭金玉[a]
(沈阳化工大学 a. 信息工程学院; b. 计算机科学与技术学院, 沈阳 110142)

摘 要: 针对一类仅使用大批历史数据结构未知的非线性工业过程,根据数据驱动及局部建模的基本思想,提出一种基于局部模型算法的在线多模型辨识策略。从向量相似的角度提出了一种新的选择数据信息(即建模邻域的确定)的方法,有效提高了获得当前时刻系统最佳局部模型的数据精确度。给出了权值选定的适合度标准及带宽 h 选择的快速方法。最后对算法进行了特性分析及仿真研究,并与其他局部建模算法的计算结果进行比较,验证了本文辨识算法的有效性。

关键词: 多模型; 向量; 邻域; 非线性系统
中图分类号: TP273 **文献标志码**: A **文章编号**: 1001-3695(2012)05-1631-03
doi:10.3969/j.issn.1001-3695.2012.05.007

Local multi-model method based on similarity of vector

ZENG Jing[a], WANG Jun[b], GUO Jin-yu[a]
(*a. College of Information Engineering, b. College of Computer Science & Technology, Shenyang University of Chemical Technology, Shenyang* 110142, *China*)

Abstract: This paper proposed an on-line multiple models identification strategy based on data-driven and local-modeling for a class of unknown-structure nonlinear systems on the basis of numerous historical database. And it proposed a new approach to determine data information (determine the neighborhood), it could improve precision of the data which used to get optimal local model. Then gave the goodness of fit criterion and rapid selection method to determine bandwidth. Finally, it did some characteristic analysis and simulation studies, and the results illustrate the validity of this approach.
Key words: multi-model; vector; neighborhood; nonlinear system

0 引言

相对于传统的全局建模方法而言,局部模型辨识在对非线性系统辨识上有明显的优势。当观察值的数量变得相当大时,确定模型结构及相关最优问题都会变得很复杂,这时全局方法便不太适合。受局部建模思想和数据库技术的鼓舞,Stenman[1]提出了一种 model-on-demand(MOD)建模策略,即假设所有历史数据都存于一个数据库中,根据实际情况的需要,从数据库中取出有用的相关数据,在线建立系统的当前模型。这便是按需求建模,是一种局部建模方法。

局部建模的思想并不是新提出来的。在统计领域中,不同学者已经对局部建模算法研究了相当长一段时间并产生了重大影响[2~5],对局部建模算法的研究是一个开放的、活跃的领域,研究和探索基于局部建模的非线性系统辨识方法,具有重大的理论价值和实际意义。

在局部建模过程中,如何确定工作点邻域 $\Omega_k(x)$,是决定模型精确与否的重要因素,也是有待深入研究的问题。工作点邻域的确定具体是指在估计当前工作点的相应输出时,从大量输入输出历史数据中找出与该工作点相近或相似的模态。即根据某种规则划定该工作点的邻域,再对该邻域内的模态建立局部模型,并估计出当前工作点的相应输出。故建模中并不事先规定各局部输入区间,而是根据当前工作点在线划定其所属的区间(邻域)。

常用的选择当前状态 x 工作点邻域 $\Omega_k(x)$ 的方法有 K-nearest neighbors(KNN)[6]、K-surrounding neighbors(KSN)[7]及 K-bipartite neighbors(KBN)方法等[8]。这些方法在解决具体问题时,获得了不错的数据选择效果。但实际上,这些基于欧式距离的信息选择准则,仅仅只考虑了数据间的距离,并没有充分挖掘数据系统的信息。本文就是依据局部建模的基本思想,提出一种基于向量相似度的局部建模方法。在给出了向量相似度定义的基础上,从向量相似的角度提出了一种新的选择数据信息(即建模邻域 $\Omega_k(x)$ 的确定)的方法,大大提高了所选数据的准确度。最后对算法进行了仿真研究,并与其他局部建模算法的计算结果进行比较,说明算法的有效性。

1 局部建模方法

假设已经存在一个足够完善的输入输出数据集合 $\{(Y_i, X_i)\}_{i=1}^N$,它可以表征非线性过程中各种可能出现的基本工况条件,过程的输入输出关系可以表示为[1]

$$Y_i = m(X_i) + \varepsilon_i \quad i = 1, \cdots, N \tag{1}$$

需要计算的是该非线性映射中某一固定操作点 x 处所对应的输出预测值。这种问题可以利用局部建模策略进行解决。

收稿日期: 2011-10-16; **修回日期**: 2011-11-30 **基金项目**: 国家自然科学基金资助项目(60874057); 辽宁省博士启动基金资助项目(20091061); 辽宁省教育厅科研项目(L2011064)

作者简介: 曾静(1981-),女,河南信阳人,副教授,博士,主要研究方向为非线性系统建模、预测控制(zengjing0066@sohu.com);王军(1978-),男,辽宁大连人,副教授,博士,主要研究方向为计算机网络;郭金玉(1975-),女,山东高唐人,副教授,博士,主要研究方向为生物特征识别、故障诊断.

设在输入向量 x 处，系统对应的局部函数关系 $m(\cdot)$ 可以用 p 阶多项式表示为

$$m(x,\beta)=\beta_0+\beta_1(X_i-x)+\cdots+\beta_p(X_i-x)^p \tag{2}$$

在获得相似数据样本后，局部建模方法的实质是一个加权最优问题，它的目的是最小化模型和数据之间的不匹配度：

$$\hat{\beta}=\arg\min_{\beta}\sum_{i\in\Omega_k(x)}e(Y_i-\sum_{j=0}^{p}\beta_j(X_i-x)^j)w_i(x) \tag{3}$$

其中：$w_i(x)$ 表示权值；$\Omega_k(x)$ 表示 x 的一个邻域，它包含着 k 个采样点。该邻域的大小由以下函数关系决定：

$$\Omega_k(x)\triangleq\{i_1,\cdots,i_k\}=\{i:d(X_i,x)\leqslant h\} \tag{4}$$

其中：$d(\cdot,\cdot)$ 是距离函数；参数 h 是带宽，它决定邻域的幅度。

2 工作点邻域的确定

在定义向量相似度 S 之前，介绍一下相关知识。

设存在两个向量 $X=(x_1,x_2,\cdots,x_n)$，$Y=(y_1,y_2,\cdots,y_n)$，则：

向量的内积为

$$[X,Y]=x_1y_1+x_2y_2+\cdots+x_ny_n \tag{5}$$

向量的范数为

$$\|X\|=\sqrt{[X,X]}=\sqrt{x_1^2+x_2^2+\cdots+x_n^2} \tag{6}$$

向量间的夹角为

$$\theta=\arccos\frac{[X,Y]}{\|X\|\cdot\|Y\|} \tag{7}$$

向量的正交：当 $\theta=90°$ 时（即 $[X,Y]=0$），称向量 X,Y 为正交向量。

由于向量包括方向和大小两个要素，故可用方向和大小来综合表征两向量的相似度。现定义如下：

定义 1 设 $X=(x_1,x_2,\cdots,x_n)$ 为参考向量，$Y=(y_1,y_2,\cdots,y_n)$ 为比较向量。向量 X 和 Y 的范数相似度 α 为

$$\alpha=\begin{cases}1-\dfrac{|\|X\|-\|Y\||}{\|X\|} & \|Y\|\leqslant 2\|X\|\\ 0 & \|Y\|>2\|X\|\end{cases} \tag{8}$$

定义 2 设 $X=(x_1,x_2,\cdots,x_n)$ 为参考向量，$Y=(y_1,y_2,\cdots,y_n)$ 为比较向量。向量 X 与 Y 的方向相似度 β 为

$$\beta=1-\frac{\theta}{90°} \tag{9}$$

由定义可知：

a）$\alpha\in[0,1]$。当 $\|Y\|\leqslant\|X\|$ 时，$\alpha=\dfrac{\|Y\|}{\|X\|}$；当 $\|X\|<\|Y\|\leqslant 2\|X\|$ 时，$\alpha=1-\dfrac{\|Y\|-\|X\|}{\|X\|}$；当 $\|Y\|\geqslant 2\|X\|$ 时，$\alpha=0$。

b）$\beta\in[-1,1]$。当 $0\leqslant\theta\leqslant 90°$ 时，$\beta\in[0,1]$；当 $90°\leqslant\theta\leqslant 180°$ 时，$\beta\in[-1,0]$。

基于以上定义和分析，本文给出了基于向量相似度方法构造输入向量的邻域准则，即在构造当前工作点 x 的邻域 $\Omega_k(x)$ 时，综合考虑向量间范数相似度 α 和方向相似度 β，给出向量相似度的定义如下：

定义 3 设 X_i 为数据库 $\{(Y_i,X_i)\}_{i=1}^{N}$ 内输入向量，x 为当前工作点。向量 X_i 和 x 的向量相似度 $S(X_i,x)$ 为向量范数相似度 α 与向量方向相似度 β 的乘积，即

$$S(X_i,x)=\alpha\cdot\beta \tag{10}$$

向量相似度 $S(X_i,x)$ 具有以下性质：

a）$S(X_i,x)\in[-1,1]$。当 $0\leqslant\theta\leqslant 90°$ 时，$S(X_i,x)\in[0,1]$；当 $90°\leqslant\theta\leqslant 180°$ 时，$S(X_i,x)\in[-1,0]$。

b）正交向量（$\theta=90°$）的相似度 $S(X_i,x)=0$。范数相同的两向量若夹角 $\theta=0$，则 $S(X_i,x)=1$；范数相同的两向量若夹角 $\theta=180°$，则 $S(X_i,x)=-1$。

由此可见，向量相似度 $S(X_i,x)$ 综合考虑了信息向量的范数和夹角信息，直接反映了向量间的相似程度。参数 α 是随向量间范数差距的减小而增大，且参数 β 也是随向量间夹角 θ 的减小而增大的。因此，两个信息向量越相似，则范数差距越小，α 越大，且夹角越小，β 也越大，从而整个 $S(X_i,x)$ 也越大。

由于 $S(X_i,x)\in[-1,0]$ 时，代表比较向量与当前工作点参考向量间夹角为 $90°\leqslant\theta\leqslant 180°$，认为两向量间夹角过大，此比较向量偏离当前工作点，不利于系统局部建模，放弃选用此信息构造建模邻域。

当 $S(X_i,x)\in[0,1]$ 时，满足

$$S(X_i,x)>h\quad S(X_i,x)\in[0,1] \tag{11}$$

输入向量即选为工作点邻域 $\Omega_k(x)$ 内的数据。其中 h 为带宽参数时。

由于带宽参数 $h>0$，故式（11）可简化为

$$S(X_i,x)>h$$

综上，给出工作点邻域选择算法：

a）找出与当前工作点最近的类，取该类所有输入向量记做输入集 K_x；给出初始样本 $i=1$。

b）取 K_x 中第 i 个样本，根据式（8）和（9）计算其与当前工作点间的范数相似度 α 及方向相似度 β。

c）由式（10）计算向量相似度 $S(X_i,x)$。若 $S(X_i,x)>h$，则保存该样本于 $\Omega_k(x)$，否则舍弃该样本。

d）$i=i+1$，转 b）。$i>\max($样本数$(K_x))$，终止。

3 权值的选择

本文选择权值的方法是基于局部多项式技术，其具体过程可分解为两部分：

a）将局部邻域内的数据与 x 之间的关系转换为距离：

$$d(X_i,x)=\|X_i-x\|_M \tag{12}$$

其中：$\|\cdot\|_M$ 表示一个尺度向量形式，距离函数的选择非常重要。

b）将距离转换为加权：

$$W_i(x)=K(\frac{d(X_i,x)}{h}) \tag{13}$$

其中：$K(\cdot)$ 是一个 Kernel 函数；h 代表带宽。要想决定权值，必须先确定距离函数、Kernel 函数及带宽的大小。

距离函数完全由度量矩阵 M 所决定，本文采用的度量函数为对角欧几里德距离：

$$d(X_i,x)=\|X_i-x\|_M=\sqrt{(X_i-x)^{\mathrm{T}}M(X_i-x)}$$
$$M=\mathrm{diag}(m_1,m_2,\cdots,m_d)$$

一般使度量矩阵 M 与回归量的逆协方差成正比。

$K(\cdot)$ 是一个 Kernel 函数，它将距离函数转换为权值。

$$K_h(\cdot)\triangleq h^{-1}K(\frac{\cdot}{h}) \tag{14}$$

其中：$K(x)=\dfrac{71}{80}(1-|x|^3)^3$（即 Tricube Kernel 函数）；$h$ 是带宽参数，它的选择对估计器性能至关重要，因为它控制着偏差与方差之间的平衡。带宽参数在本质上决定着邻域的大小。

计算带宽 h 时所采用的适合度标准有许多种,如局部 cross-validation 标准:

$$\mathrm{CV}(\boldsymbol{x},k)=\frac{1}{tr(\boldsymbol{W}_k)}\sum_{i\in\Omega_k(\boldsymbol{x})}w_i(\boldsymbol{x})\left(\frac{Y_i-\bar{m}(\boldsymbol{X}_i,k)}{1-\inf l(\boldsymbol{x},\boldsymbol{X}_i)}\right)^2$$

局部 generalized cross-validation 标准:

$$\mathrm{GCV}(\boldsymbol{x},k)\triangleq tr(\boldsymbol{W}_k)\frac{\sum_{i\in\Omega_k}w_i(\boldsymbol{x})(Y_i-\bar{m}(\boldsymbol{X}_i,k))^2}{tr(\boldsymbol{W}_k)-tr(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{W}_k\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{W}_k^2\boldsymbol{X}_k)^2}$$

其中:$\bar{m}(\boldsymbol{X}_i,k)\triangleq\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{X}_i-\boldsymbol{x})\hat{\beta}$。

还可采用局部 AIC、FPE、Mallows c^p 标准等,它们的目的都是最小化 MSE 或其他的类似标准。采用上述适合度标准确定最优带宽的方法一般为:在已选定的输入输出数据集合中,首先选定一个最小邻域尺度,在此基础上逐渐增加邻域的大小,针对不同的邻域尺度分别计算模型参数估计值并利用适合度标准进行评估;当适合度标准值达到最小时对应的带宽取为最优带宽,它反映了方差与偏差之间一个好的平衡,它对应的邻域尺度为最终选择的邻域尺度,它对应的模型参数估计值为最优估计值。

以上所提到的适应算法要求逐步增加带宽的大小,即增加邻域尺度,并对不同的带宽计算适合度标准,只有当适合度标准值最小时对应的带宽才是最优带宽。而重复这种计算将是一种很复杂且繁琐的工作。为了加速带宽的选择速度,本文采用下面的方法:

a)给定一个小带宽 h_0,使它接近于能获取估计值的最小带宽即可。指数级增加带宽:

$$h_i=C_h h_{i-1}\ (C_h>1)$$

b)针对不同的带宽计算模型参数估计值及适应度标准值。

c)适合度标准值最低时对应的带宽便为最优带宽 h_{opt}。

通常,$C_h=1+\frac{0.3}{d}$。d 表示回归空间的维数。

在 Kernel 函数、带宽参数 h、距离函数 d 等都选定后,便可以利用式(13)计算权值了。

4　多模型辨识算法小结

综上可知,本文提出的非线性系统多模型局部建模算法实现的具体步骤如下:

a)确定输入输出历史数据库 $\{(Y_i,\boldsymbol{X}_i)\}_{i=1}^N$,并对数据信息进行必要的预处理。

b)确定与当前工作点最近的类,得出当前工况下对应输入向量 $\boldsymbol{x}$ 的具体工作点邻域 $\Omega_k(\boldsymbol{x})$。

c)逐步增加邻域 $\Omega(k)$ 的大小,利用式(3)计算 $\hat{\beta}$ 及相应的适合度标准,直至标准值降至最低或已经达到了最大邻域范围。

d)通过存储在 c)中的结果,找出标准值最低时所对应的最优带宽 h_{opt}。

e)由最优带宽 h_{opt} 得出最优参数估计值 $\hat{\beta}^{(h_{\mathrm{opt}})}$,并计算出当前输出估计值 $\hat{m}(x)$,噪声方差 $\hat{\sigma}^2(x)$ 及输出估计值方差 $\mathrm{var}(\hat{m}(x))=\hat{\sigma}^2(x)\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{W}_k$。

在得到 x 点的输出估计值后,评定的性能指标可采用平均平方误差:

$$\mathrm{MSE}(\hat{m}(x,h))\triangleq E(\hat{m}(x,h)-m(x))^2 \qquad (15)$$

5　仿真分析

考虑以下非线性系统(Narendra & Li, 1996):

$$x_1(t+1)=\left(\frac{x_1(t)}{1+x_1^2(t)}+1\right)\sin(x_2(t))$$

$$x_2(t+1)=x_2(t)\cos(x_2(t))+x_1(t)\exp\left(-\frac{x_1^2(t)+x_2^2(t)}{8}\right)+\frac{u^3(t)}{1+u^2(t)+0.5\cos(x_1(t)+x_2(t))}$$

$$y(t)=\frac{x_1(t)}{1+0.5\sin(x_2(t))}+\frac{x_2(t)}{1+0.5\sin(x_2(t))}+e(t)$$

其中,$x_1(t)$ 与 $x_2(t)$ 为系统的状态变量;$y(t)$,$u(t)$ 与 $e(t)$ 分别为系统的输出变量、输入变量和白噪声。输入输出数据库由 3 000组样本组成,同时在样本内加入了随机扰动;另外产生 300 组测试数据用于模型校验。产生的 3 000 组输入输出数据库如图 1 所示。

利用本文所提出的基于局部模型的多模型算法进行模型辨识。这里输入输出模型结构采用 ARX 模型结构,在 ARX 模型结构集($1\leqslant n_a\leqslant 5, 1\leqslant n_b\leqslant 5, n_k=1$)中选取合适的模型结构,最后确定为 ARX331 模型结构,即

$$\hat{y}(t)=m\begin{pmatrix}y(t-1),y(t-2),y(t-3)\\u(t-1),u(t-2),u(t-3)\end{pmatrix}$$

模型校验时采用的输入信号为

$$u(t)=\sin\left(\frac{2\pi t}{10}\right)+\sin\left(\frac{2\pi t}{25}\right)$$

仿真结果如图 2 所示,其中实线为实际输出,虚线为估计输出。可以看出,采用本文算法进行局部多模型辨识,可以很好地逼近非线性系统的真实输出。为了说明本文算法的优越性,与 MOD[9] 和 lazy learning[10] 方法对该问题的辨识结果进行比较,比较结果如表 1 所示。

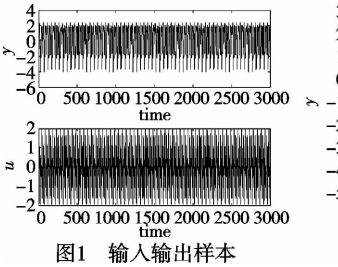

图1　输入输出样本

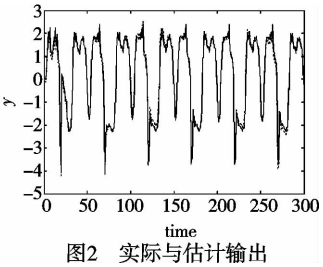

图2　实际与估计输出

表 1　各种算法结果比较

算法	max E	MSE
MOD	0.490 3	0.126 3
lazy learning	0.45 10	0.129 9
本文算法	0.298 2	0.100 0

6　结束语

为了使系统能自适应地反映实际过程中工况变化的特性,本文考虑采用基于数据驱动的方法对非线性系统进行局部建模,从向量相似的角度提出了一种新的选择数据信息(即建模邻域的确定)的方法,既考虑了数据向量间的范数相似度,又考虑了数据向量间的角度相似度,有效提高了获得当前时刻系统最佳局部模型的数据精确度。对该算法进行了特性分析,并通过仿真对比说明了本文改进算法的有效性。

参考文献:

[1] STENMAN A. Model-on-demand: algorithms, analysis and applications[D]. Linkoping, Sweden: Linkoping University, 1999.

[2] 邹涛,王昕,李少远. 基于混合逻辑的非线性系统多模型预测控制[J]. 自动化学报,2007,33(2):188-192.　(下转第 1640 页)

(上接第 1633 页)

[3] LI Ning, LI Shao-yuan, XI Yu-geng. Multiple model predictive control for MIMO systems[J]. Acta Automatica Sinica, 2003, 29(4): 516-523.

[4] ROLL J, BEMPORAD A, LJUNG L. Identification of piecewise affine systems via mixed-integer programming[J]. Automatica, 2004, 40(1): 37-50.

[5] ROLL J, NAZIN A, LJUNG L. Nonlinear system identification via direct weight optimization[J]. Automatica, 2005, 41(3): 475-490.

[6] LIU Hua-wen, ZHANG Shi-chao, ZHAO Jian-ming, *et al*. A new classification algorithm using mutual nearest neighbors[C]//Proc of the 9th International Conference on Grid and Cloud Computing. Washington DC: IEEE Computer Society, 2010: 52-57.

[7] ZHANG Jian-ping, YIM Y S, YANG Jun-ming. Intelligent selection of instances for prediction funcfions in lazy learning algorithms[J]. Artificial Intelligence Review, 1997, 11(1-5): 175-191.

[8] LI Q, MITIANOUDIS N, STATHAKI T. Spatial kernel K-harmonic means clustering for multi-spectral image segmentation[J]. IET Image Processing, 2007, 1(2): 156-167.

[9] LEE H, RIVERA D E, MITTELMANN H D, *et al*. Optimization-based design of plant-friendly input signals for model-on-demand estimation and model predictive control[C]//Proc of American Control Conference. [S. l.]: IEEE Press, 2007: 1560-1565.

[10] KOBAYASHI M, KONISHI Y, ISHIGAKI H. A lazy learning control method using support vector regression[C]//Proc of Mediterranean Conference on Control & Automation. 2007: 1-7.