文章编号: 1001-0920 (2011) 10-1459-04

没有参数的系统辨识方法与模型估计方式

刘开第, 庞彦军, 马丽涛

(河北工程大学 不确定数学研究所，河北 邯郸 056038)

摘 要: 基于模型类/参数模式的传统系统辨识, 虽然囊括了系统辨识几乎所有的成果, 理论也趋于成熟, 但不宜使用在空间中分布不均匀且数量相对少的数据. 鉴于此, 提出针对这类数据建模的无参数系统辨识研究方向, 讨论基于代表点和加权距离的无参数系统辨识方法, 给出基于分类一致性准则的模型估计方式. 与传统系统辨识的区别是, “没有参数”并且从实质上改变估计模型的方式. 用 IRIS, Breast Cancer 等典型数据检验了模型的有效性.

关键词: 系统辨识；无参数辨识；指标分类权；分类一致性；学习算法

中图分类号: TP13 **文献标识码:** A

Method of nonparametric system identification and model estimation

LIU Kai-di, *PANG Yan-jun*, *MA Li-tao*

(Institution of Uncertainty Mathematics，Hebei University of Engineering，Handan 056038，China．Correspondent: LIU Kai-di，E-mail：liukaidi@hebeu.edu.cn)

Abstract: Although traditional system identification methods based on model class/parameter mode include almost all the results of system identification and the theory is also mature, it is not appropriate to the relatively less data which is uneven distributed in space. For modeling such data, nonparametric system identification research field is found. Based on representative points and the weighted distance, this paper discusses nonparametric system identification, and proposes model estimation method based on classification “consistency” criterion as well. Finally, with typical data such as IRIS and Breast Cancer data, the effectiveness of model is verified.

Key words: system identification；nonparametric identification；index classification weight；classification consistency; learning algorithm

1 引 言

Zadeh[1]于 1962 年首次提出了系统辨识的概念, Liung[2]于 1978 年将模型类、数据与准则作为系统辨识三要素, 并将系统辨识定义为: 辨识就是按着一个准则在一组模型类中选择一个与数据拟合得最好的模型. 由于 Liung 给出的定义更加实用、便于操作, 所以系统辨识 (称为传统系统辨识) 得以迅速发展. 到 20 世纪 90 年代末, 传统系统辨识的理论已趋于成熟, 各种不同的非线性辨识模型与方法大量涌现. 如基于神经网络的系统辨识[3-5]、基于小波网络的系统辨识[6-8]、基于模型逻辑的系统辨识[9-10]、T-S 模糊模型[11-13]、基于遗传算法的系统辨识[14-16]、基于支持向量机 (SVM) 的系统辨识[17-18]以及融合模糊逻辑、神经网络和遗传算法的系统辨识[19-20]等. 这些不同的非线性模型与方法广泛应用于基于数据的复杂系统的控制、决策、调度和故障诊断中.

虽然基于模型类/参数模式的传统系统辨识囊括了当今系统辨识几乎所有成果, 理论也趋于成熟, 但模型类/参数并不是系统辨识的惟一模式, 传统系统辨识也不能涵盖系统辨识的全部理论、内容与方法. 理由是: 1) 任何基于数据的实际系统, 理想模型一定是针对数据的“个性化”模型, 但“模型类/参数”模式因为极具“通用性”, 不利于凸显数据的“个性化”特点, 使得从模型类出发建模的传统系统辨识难以获得针对数据的“个性化”模型. 2) 传统系统辨识的理论基础是统计学习理论, 要求数据尽可能充满特征空间, 数据个数应趋于无穷大, 以保证有足够数据供学习使用. 由于实际系统获取的数据经常是有限的, 不能充满特征空间, 也不能体现更多的统计规律, 此时, 传统系统辨识方法无能为力. 何种系统辨识理论与方法能够

收稿日期: 2010-05-18；**修回日期:** 2010-08-11.

基金项目: 国家自然科学基金项目(60874116, 60940036)；河北省自然科学基金项目(F2009000857).

作者简介: 刘开第(1940−), 男, 教授, 从事不确定信息数学处理方法等研究; 庞彦军(1964−), 男, 教授, 从事不确定信息数学处理方法等研究.

根据这样的数据获取到针对数据的“个性化”模型呢? 实际上, 获取针对数据的“个性化”模型的最直观理由是直接从数据出发建模, 为此提出不设模型类、直接从数据出发建模的无参数系统辨识研究方向.

因为不设模型类, 辨识过程没有参数; 由于不涉及基于参数的“误差估计函数”, 可从实质上改变估计模型的方式. 这正是Liung设想中所期望的[21]: “我们真正感兴趣的是, 学习理论能否帮助我们创建一种没有参数的系统辨识理论, 是否能从实质上改变我们估计模型的方式”. 无论Liung设想能否实现, 至少在Liung看来, “模型类/参数”并不是系统辨识的惟一模式; 基于参数的“误差估计函数”也不应是估计模型的惟一方式. 但问题是, 怎样才能构建这种“从实质上改变我们估计模型方式的无参数”系统辨识模型呢?

无参数系统辨识的依据是, 同类数据在空间中是“相互接近的”. 无参数系统辨识策略是, 依据各同类数据所在的空间小区域, 构建一种数学模型, 这种数学模型有两个功能:

1) 对于空间中任意样本点具有分类功能, 即可识别该样本属于哪一个数据类.

2) 模型对训练样本的分类应“尽可能”保持分类“一致性”, 即将模型对训练样本分类时出现的分类“不一致”样本个数控制在允许范围内.

无参数系统辨识的实施方法:

1) 如果同类数据在空间中大致呈“球形”分布, 用一个代表点 m_k 近似代表 C_k 数据类; 用样本点 x 到代表点 $m_k(k=1,2,\cdots,p)$ 的“某种距离” $d(x,m_k)$ 作为样本点 x 与 C_k 类间的“相似性度量”. 在“最小距离识别准则”下, 可得到一个数学模型 $A(x,m_k,d(x,m_k))$, 该模型显然对于空间中任意样本点 x 具有分类功能. 由于要求模型对训练样本“尽可能”保持分类“一致性”, 必须确定合适的“相似性度量” $d(x,m_k)$ 和选择满足条件的“代表点” $m_k(k=1,2,\cdots,p)$.

2) 由基于代表点的无参数系统辨识过程可知, 是否所有的同类数据在空间均大致呈“球形”分布, 不呈“球形”分布的数据称为特别复杂数据, 对于特别复杂数据可用“点集”代表类. 因为基于代表点的无参数系统辨识可为基于点集的无参数系统辨识提供必要的方法支持, 所以先研究基于代表点的无参数系统辨识.

2 相似性度量的确定方法

2.1 特征空间的无量纲化

无参数系统辨识需要在空间中定义“两点距离”, 必须对 d 维特征空间进行无量纲化变换. 令

$$x_{ij}=\frac{z_{ij}-\min\limits_i\{z_{ij}\}}{\max\limits_i\{z_{ij}\}-\min\limits_i\{z_{ij}\}},$$
$$i=1,2,\cdots,M,\ j=1,2,\cdots,d. \tag{1}$$

其中: z_{ij} 是样本 z_i 的第 j 维分量, M 是包含训练样本与待识样本在内的样本总数, d 是特征空间的维数.

变换后的 d 维空间称为标称化空间, 其中每个样本 x_i 的各维分量均在闭区间 [0,1] 上.

2.2 空间知识提取与指标分类权

当用代表点 $m_k(k=1,2,\cdots,p)$ 代表 C_k 数据类时, 需要知道各维空间指标发生了怎样的“不对称性”变化, 因为这种变化反映了分类数据的特点, 也影响着样本分类. 令

$$m_k=(m_{k_1},m_{k_2},\cdots,m_{k_d}),\ k=1,2,\cdots,p; \tag{2}$$

$$\bar{m}=\frac{1}{p}\sum_{k=1}^{p}m_k=(\bar{m}_1,\bar{m}_2,\cdots,\bar{m}_d),$$
$$k=1,2,\cdots,p; \tag{3}$$

$$\sigma_j^2=\frac{1}{p}\sum_{j=1}^{p}(m_{kj}-\bar{m}_j)^2,\ j=1,2,\cdots,d; \tag{4}$$

$$\lambda_j=\sigma_j\Big/\sum_{t=1}^{d}\sigma_t,\ j=1,2,\cdots,d. \tag{5}$$

称 λ_j 为用代表点 $m_k(k=1,2,\cdots,p)$ 代表 C_k 时 j 指标的分类权, 显然分类权满足

$$0\leqslant\lambda_j\leqslant 1,\ \sum_{j=1}^{d}\lambda_j=1. \tag{6}$$

分类权 λ_j 的作用如下:

1) 如果某个 $\lambda_j=0$, 此时 p 个代表点的 j 分量均相等, 说明 j 指标对于将 p 个代表点区分开不起作用, 即 j 对当前分类不起作用.

2) 若 $\lambda_j>0$, 则 λ_j 越大时 σ_j 越大, 此时从 j 轴上看, p 个代表点越分散, 说明 j 指标对于将 p 个代表点区分开贡献大, 即对当前分类贡献大.

2.3 加权距离与相似性度量

用代表点 $m_k(k=1,2,\cdots,p)$ 代表 C_k 类, $d(x,m_k)$ 是样本点 x 与代表点 m_k 间的某种距离. 如果将 $d(x,m_k)$ 作为 x 与 C_k 类的相似性度量, 则 $d(x,m_k)$ 因包含了样本 x 的分类信息而不再是单纯距离. 对于分类而言, 各维空间指标的地位是不对称的, 当 $\lambda_j=0$ 时, j 指标对当前分类不起作用, 指标 j 的分量自然不能用于计算相似性度量. 所以, $d(x,m_k)$ 应是以 λ_j 为权的一种加权距离, 即

$$[d(x,m_k)]^2=\sum_{j=1}^{d}\lambda_j(x_j-m_{kj})^2,\ k=1,2,\cdots,p. \tag{7}$$

其中: x_j 是样本 x 的 j 分量, λ_j 是 j 指标的分类权.

因为不同数据对应不同的分类权, 所以用加权距离表示的相似性度量中包含着数据的“个性化”特

点. 在“最小加权距离识别准则”下的数学模型 $A(x, m_k, d(x,m_k))$ 随代表点 $m_k(k=1,2,\cdots,p)$ 的确定而确定, 该模型显然对于空间中任意样本点 x 具有分类功能. 基于代表点的无参数系统辨识归结为: 如何按着分类“一致性”准则来选择合乎要求的代表点.

3　选择代表点的学习算法

3.1　初始代表点

设标称化指标空间中 N 个训练样本被分成 p 个类, $C_k(k=1,2,\cdots,p)$ 为第 k 个数据类, 内含 N_k 个训练样本, $\sum_{k=1}^{p} N_k = N$. 选择 C_k 类的均值 $m_k(0)(k=1,2,\cdots,p)$ 作为 C_k 类的初始代表点, 由此可设计选择代表点的学习算法.

3.2　选择代表点的算法步骤

Step 1: C_k 类的类均值 $\bar{m}_k(k=1,2,\cdots,p)$ 作为 C_k 的初始代表点, 记为 $m_k(0)$.

Step 2: 迭代算法按节拍 t 进行, 置 $t=1$, 最大迭代次数 $t_{\max}$, 终止常数 $\varepsilon>0$.

Step 3: 计算当前代表上各指标的分类权 $\lambda_j(t-1),\ j=1,2,\cdots,d$.

Step 4: 计算训练样本点 $x_i(i=1,2,\cdots,N)$ 到各类代表点 $m_k(t-1)$ 的加权距离为

$$[d_{ik}(t-1)]^2=\sum_{j=1}^{d}\lambda_j(t-1)[x_{ij}-m_{kj}(t-1)]^2,$$

$$i=1,2,\cdots,N,\ k=1,2,\cdots,p.$$

最小距离为 $d_{ik_i}(t-1),\ i=1,2,\cdots,N$, 其中 k_i 是与最小加权距离对应的代表点 (x_i 的最小点) 的序号.

Step 5: 若 $x_i\in C_k$ 且 $k_i\neq k$, 则 x_i 是 C_k 类中被错分到 C_{k_i} 类的分类“不一致”训练点. 在 N 个训练点中, 假设有 M_t 个分类“不一致”点, 依次记为 $x_1(t-1)$, $x_2(t-1),\cdots,x_{M_t}(t-1)$.

Step 6: 若 $M_t=0$ (无不一致点), 则转至 Step 12; 否则继续.

Step 7: 令 $m_k^{(q)}(t-1)=m_k(t-1),\ k=1,2,\cdots,p,\ q=1$.

Step 8: 按下式用不一致点修正当前的各代表点:

$$m_k^{(q+1)}(t-1)=\begin{cases} m_j^{(q)}(t-1)+\dfrac{\alpha(0)}{t}[x_q(t-1)-m_j^{(q)}(t-1)], \\ \qquad x_q(t-1)\in C_j,\ k=j; \\ m_{k_q}^{(q)}(t-1)-\dfrac{\alpha(0)}{t}[x_q(t-1)-m_{k_q}^{(q)}(t-1)], \\ \qquad x_q(t-1)\in C_j,\ k=k_q; \\ m_k^{(q)}(t-1),\ x_q(t-1)\in C_j, k\neq j, k\neq k_q. \end{cases} \tag{8}$$

其中 $\alpha(0)>0$ 为控制常数.

Step 9: 若 $q<M_t$, 则令 $q=q+1$, 返回 Step 8; 否则继续.

Step 10: 令 $m_k(t)=m_k^{(M_t+1)}(t-1)$, 若

$$\sum_{k=1}^{p}\sum_{j=1}^{d}\lambda(t-1)[m_{k_j}(t)-m_{k_j}(t-1)]^2<\varepsilon,$$

则转至 Step 12; 否则继续.

Step 11: 若 $t<t_{\max}$, 则令 $t=t+1$, 返回 Step 3; 否则继续.

Step 12: 迭代停止, 输出当前代表点 $m_1(t), m_2(t), \cdots, m_p(t)$, 并由 Step 3 计算当前代表点条件下的分类权 $\lambda_1(t), \lambda_2(t), \cdots, \lambda_d(t)$.

3.3　数据的复杂性分类

用当前代表点 $m_1(t), m_2(t), \cdots, m_p(t)$ 依次代表 $C_1\sim C_p$ 数据类, 按最小加权距离识别准则生成的数学模型, 可对训练样本进行分类. 若不出现分类“不一致”点, 则称分类数据为简单数据, 对于简单数据, 判待识点 y 属于其最小点代表的类; 若每类中“不一致”点个数都相对较少, 则称数据为较复杂数据. 对于较复杂数据, 若 x^* 是 C_k 类的一个不一致点, 则设置 x^* 的一个 δ 邻域, 当待识点 y 落入某个 x^* 的 δ 邻域内, 判断 y 与 x^* 同类, 对于落在 δ 邻域以外的待识点 y, 则判断 y 属于其最小点代表的类; 若至少有一类出现相对多的“不一致”点, 则称数据为复杂数据, 对于复杂数据, 需用一个“点集”代表类.

4　模型有效性检验

1) IRIS 数据检验

IRIS 数据[22]是国际上公认的检验分类效果的检验数据. IRIS 数据分 3 类, 每类 50 个样本, 每个样本都是关于花瓣测量值的 4 维数据. 实验前, 先按式 (1) 对数据施行无量纲化变换.

实验 1　目的: 检验加权距离与欧氏距离作为样本与类之间相似性度量时的分类效果.

作法: 3 类 IRIS 数据, 每类 50 个样本, 用每类的类均值作为该类代表点, 分别以加权距离和欧氏距离作为样本与类之间的相似性度量, 对全部 150 个样本重新分类, 比较两种相似性度量下分类“不一致”的样本个数.

结果: 欧氏距离下, 不一致样本 10 个; 加权距离下, “不一致”样本 6 个. 加权距离的分类效果优于欧氏距离, 说明从数据中提取空间知识信息作为确定样本与类之间相似性度量的启发性知识, 有利于增加所建模型对数据的针对性.

实验 2　目的: 检验模型对训练样本分类的有效性.

作法: 对于 3 类共 150 个数据, 用类均值作初始

代表点, 取步长$\alpha(0)=0.17$, 经迭代算法后选择代表点. 对150个样本按最小加权距离重新分类.

结果: 经多次实验, 不一致样本数稳定在2个, 个别时候为1个. 经学习算法后选择代表点, “不一致”样本数由6个稳定在不超过2个, 对于训练样本的正确识别率超过98%, 说明模型十分有效, 也说明按本文方法IRIS数据属于较复杂数据.

实验3 目的: 检验模型的外延有效性.

作法: 从每类50个样本中, 随机选出25个, 3类共75个作为训练样本, 其余75个作为检验样本. 以均值为初始代表点, 取步长$\alpha(0)=0.17$, 经迭代算法后确定代表点, 分别对75个训练样本和75个检验样本分类, 分别计算分类“不一致”点, 重复100次实验.

结果: 100次实验, 对训练样本的分类“不一致”样本个数均为0个. 100次实验, 对检验样本的分类“不一致”样本个数为2∼3个, 平均2.58个, 正确识别率超过96%, 说明模型外延性良好.

2) Breast Cancer数据检验

Breast Cancer数据原有699组, 每维均是9维数据, 分为benign和maligant两类, 剔除缺省的16组还余683组, 其中444组属于benign, 239组属于maligant. 实验前先用式(1)对数据无量纲化.

从444组benign类数据中随机选择222组, 从239组maligant类数据中随机选择120组, 共计342组数据作为训练样本集; 剩余341组为检验数据.

以各类训练数据的类均值作为初始代表点, 取$\alpha(0)=0.04$, 选择出当前最优代表点$m_1(t)$和$m_2(t)$, 按最小加权距离识别准则, 对检验集计算分类不一致样本数. 重复实验100次.

实验结果: 在341组检验数据中平均错分9.25组, 对检验样本的正确分辨率超过97%, 说明无参数系统辨识模型的有效性令人满意.

两种不同检验数据的检验结果表明, 无参数系统辨识模型的有效性令人满意, 主要原因是: 1)给定的同类数据在空间中大致呈“球形”分布, 所以适合用基于代表点的无参数系统辨识; 2)无参数系统辨识模型较充分地利用了数据的个性化特点.

5 结 论

空间中分布不均匀、数量相对少、无更多统计规律的数据对于实际系统而言是大量存在的, 这种数据的“个性化”特点强烈, 并且基于模型类/参数模式的传统系统辨识不再适用, 即使使用也难获得针对数据的“个性化”模型. 鉴于此, 提出不设模型类、直接从数据出发建模的无参数系统辨识研究方向. 无参数系统辨识与传统系统辨识依据不同的模式、不同的准则, 并且各自适用不同数据类型, 因此是系统辨识的两个极具互补性的不同方向. 但人们习惯了有参数的系统辨识, 接受无参数系统辨识会遇到困难, 且无参数系统辨识刚刚起步, 需要得到业内人士的关注与支持.

参考文献(References)

[1] Zadeh L A. From circuit theory to system theory[J]. Proc IRE, 1962, 50(5): 856-865.

[2] Liung L. Convergence analysis of parametric identification methods[J]. IEEE Trans on Automatic Control, 1978, 23: 770-783.

[3] Krzyza K A. Nonparametric estimation and classification using radial basis function nets and empirical risk minimization[J]. IEEE Trans on Neural Networks, 1996, 7(2): 475-487.

[4] Narendra K S, Parthasarathy K. Identification and control of dynamical system using neural networks[J]. IEEE Trans on Neural Networks, 1990, 1(1): 4-27.

[5] 魏民祥, 闫桂荣, 沈亚鹏. 基于动态神经网络非线性结构辨识的研究[J]. 应用力学学报, 2002, 17(2): 110-114.
(Wei M X, Yan G R, Shen Y P. The study of non-linear structure's identification using dynamic neural network[J]. Chinese J of Applied Mechanics, 2002, 17(2): 110-114.)

[6] Bakshi B R, Stephanopoulos G. Wave-net: A multiresolution, hierarchical neural network with localized learning[J]. AIChE J, 1993, 39(1): 57-81.

[7] Zhang Q. Using wavelet network in nonparametric estimation[J]. IEEE Trans on Neural Network, 1997, 8(2): 227-236.

[8] 王海清, 宋执环, 李平. 采用正交小波网络的非线性系统辨识方法[J]. 控制理论与应用, 2001, 18(2): 200-204.
(Wang H Q, Song Z H, Li P. Nonlinear system identification using orthogonal wavelet networks[J]. Control Theory and Applications, 2001, 18(2): 200-204.)

[9] 陈建勤, 席裕庚. 用模糊模型在线辨识非线性系统[J]. 自动化学报, 1998, 24(1): 90-94.
(Chen J Q, Xi Y G. On-line identification of nonlinear systems using fuzzy model[J]. Acta Automatica Sinica, 1998, 24(1): 90-94.)

[10] 李人厚, 张平安. 关于模糊辨识的理论与应用实际问题[J]. 控制理论与应用, 1995, 12(2): 129-137.
(Li R H, Zhang P A. On the problems of fuzzy identification theory and its pratical applications[J]. Control Theory and Applications, 1995, 12(2): 129-137.)

[11] Johansen A, Shorten R, Murray-smith R. On the interpretation and identification of dynamic takagi-sugeno fuzzy models[J]. IEEE Trans on Fuzzy Systems, 2000, 8(3): 297-313.

(下转第1468页)