基于函数集信息量的模型选择研究

盛守照　　王道波　　王志胜　　黄向华

（南京航空航天大学自动化学院　南京　210016）

摘　要：提出了子空间信息量（SIQ）和函数集信息量（FSIQ）概念，详细讨论了基于函数集信息量的模型选择问题，给出了有限含噪声样本下模型选择的近似解决方法，很好地克服了模型选择过程中普遍存在的欠学习和过学习问题，大大提高了预测模型的泛化性能，在此基础上提出了一种可行的次优模型选择算法。最后通过具体实例验证了上述方法的可行性和优越性。

关键词：子空间信息量，函数集信息量，模型选择，统计学习理论

中图分类号：TP18　　**文献标识码**：A　　**文章编号**：1009-5896(2005)04-0552-04

Research on Model Selection Based on Function Set Information Quantity

Sheng Shou-zhao　　Wang Dao-bo　　Wang Zhi-sheng　　Huang Xiang-hua

(*College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016, *China*)

Abstract　The concepts of the Subspace Information Quantity(SIQ) and Function Set Information Quantity(FSIQ) are presented; Then the problem of model selection based on FSIQ are discussed explicitly, and the approximate method of model selection based on limited samples with white noise is proposed, which resolves the problem of underfitting and overfitting of model selection and improves the generalization of predict model well. A new suboptimal algorithm for model selection is given, and its reliability and advantage are illustrated through concrete test.

Key words　Subspace Information Quantity(SIQ), Function Set Information Quantity(FSIQ), Model selection, Statistical learning theory

1　引言

有监督学习利用已知样本数据寻找合适的预测模型逼近未知的系统输入输出关系，在这个过程中不仅要避免过学习，还要尽可能提高预测模型的泛化性能，降低模型对未学习数据的预测误差[1,2]。迄今，国内外诸多学者提出了许多有监督学习方法，但大部分在实际应用中均存在一些弊端，例如随机梯度下降和误差反向传播等方法普遍采用的经验风险最小化原则在样本有限时是不合理的；贝叶斯学习过分依赖样本的先验概率信息，但实际问题往往仅有单纯的样本信息[3,4]；统计学习理论提出结构风险最小化原则的学习策略，模型选择是通过对它的推广性界的估计进行的，但遗憾的是，目前尚没有关于如何计算任意函数集的复杂性（Vaplic-Chervonenkis维）以及推广性界的一般性理论[5]；另外，不少学者提出了其它有益的方法，在模型选择中增加对预测模型复杂性的惩罚项，从而降低预测风险，然而困难的是确定惩罚函数往往需要丰富的先验知识，没有统一的确定的方法[6-8]。

实际上，上述有监督学习可以归结为一种模型选择问题，在这过程中既要防止预测模型过于复杂所导致的过学习问题，使得预测模型包含了大量噪声信息，也要防止由于预测模型过于简单所导致的欠学习问题，这两种情况都大大降低预测模型的泛化性能[9-11]，因此寻找解决上述模型选择问题的合适方法对提高预测模型泛化性能至关重要。在这方面，主要有基于信息统计[12-16]、贝叶斯统计[3,4]和结构风险最小化[1,5]等学习方法，前面已经指出了后两种方法在实际应用中存在的困难，基于信息统计的方法以Akaike提出的Akaike's Information Criterion(AIC)准则最为典型[13]，然而AIC准则不但要求真实模型包含于待学习的函数集空间中，而且它所给出的是概率意义上渐近的无偏估计，在有限样本情况下失去意义[2,14]。

本文针对上述问题，另辟捷径，提出了子空间信息量（Subspace Information Quantity, SIQ）和函数集信息量（Function Set Information Quantity, FSIQ）概念；详细讨论了基于函数集信息量的模型选择问题，给出了有限含噪声样

2003-12-09 收到，2004-03-26 改回
航空基金（01C52015）资助课题

本下模型选择的近似解决方法，很好地克服了模型选择过程中普遍存在的欠学习和过学习问题，大大提高了预测模型的泛化性能；在此基础上提出了一种可行的次优模型选择算法。上述理论和方法具有较大的理论指导和实际应用价值，并通过具体数值试验验证了其可行性和优越性。

2　子空间信息量和函数集信息量概念

2.1 子空间信息量

设 V 是数域 P 上有限维线性空间，V_1,V_2 是 V 的子空间；$\alpha\in V$，且 $\|\alpha\|\neq 0$，分别为 α 在子空间 V_1,V_2 上的投影；$\beta_1,\cdots,\beta_s$ 是 V_1 中任意一组基；$B=[\beta_1,\cdots,\beta_s]$，$\Gamma=B\left(B^{\mathrm{T}}B\right)^{-1}B^{\mathrm{T}}$。

定义 1　定义子空间信息量 $Q(V_1,\alpha)\overset{\text{def}}{=\!=}-\ln\left(1-\dfrac{\alpha^{\mathrm{T}}\Gamma\alpha}{\|\alpha\|^2}\right)$ 为 V_1 包含 α 的信息量。

由投影定理可知，$\dfrac{\|\beta-\alpha\|^2}{\|\alpha\|^2}=1-\dfrac{\alpha^{\mathrm{T}}\Gamma\alpha}{\|\alpha\|^2}$，因此，$Q(V_1,\alpha)\in[0,\infty)$。下面给出子空间信息量的若干重要性质。

性质 1　$\lim\limits_{Q(V_1,\alpha)\to 0}\beta=0$，$\lim\limits_{Q(V_1,\alpha)\to\infty}\beta=\alpha$。

性质 2　若 $V_1\supseteq V_2$，则 $Q(V_1,\alpha)\geq Q(V_2,\alpha)$。

性质 3　$Q(V_1,\alpha)\geq Q(V_2,\alpha)$ 的充要条件：$\|\beta-\alpha\|\leq\|\gamma-\alpha\|$。

子空间信息量 $Q(V_1,\alpha)$ 为度量子空间 V_1 包含空间 V 中的向量 α 的信息量提供了尺度。

2.2　函数集信息量

设 $y=g(x)$ 为 x,y 之间确定的函数关系，其中 $x\in R^p$，$y\in R$；函数集 $\Phi=\{\phi(x,\theta),\theta\in\Theta\}$，$\theta$ 和 Θ 分别是函数 $\phi(x,\theta)$ 的参数向量及对应的定义域空间，Φ 为函数集 Φ 张成的函数空间；样本集 $Z=\{(x_i,y_i)\}_{i=1}^n$，n 为样本数；$\varphi(\theta)=\left[\phi(x_1,\theta),\cdots,\phi(x_n,\theta)\right]^{\mathrm{T}}$；$E=\{\varphi(\theta),\theta\in\Theta\}$，$V$ 为 E 生成的线性空间；$y=[y_1,\cdots,y_n]^{\mathrm{T}}$；$A_2=\dfrac{1}{n}\sum\limits_{i=1}^n y_i^2$ 为样本 $\{y_i\}_{i=1}^n$ 的二阶原点矩。

定义 2　定义函数集信息量 $Q(\Phi,Z)\overset{\text{def}}{=\!=}Q(V,y)$ 为函数集 Φ 形成的函数空间 Φ 包含样本集 Z 的信息量。

函数集信息量 $Q(\Phi,Z)$ 不仅与函数集 Φ 有关，还与样本集 Z 有关。从子空间信息量的有关性质可以看出，$Q(\Phi,Z)$ 从量上直观地反映了 $g(x)$ 相对于样本点来说在函数集 Φ 上的最佳逼近程度。

定理 1　$\exists h(x)\in\Phi$，使得 $\int\left[g(x)-h(x)\right]^2\mathrm{d}x=0$ 的充要条件：$\lim\limits_{n\to\infty}Q(\Phi,Z)\to\infty$。

定理 2　$\exists h(x)\in\Phi$，对于 $\forall\eta>0$，使得 $\dfrac{1}{n}\sum\limits_{i=1}^n\left(h(x_i)-y_i\right)^2\leq\eta$ 的充要条件：$Q(\Phi,Z)\geq\ln\dfrac{A_2}{\eta}$。

定理 3　设 $\eta=A_2\exp\left(-Q(\Phi,Z)\right)$，则

$$\min_{h(x)\in\Phi}\frac{1}{n}\sum_{i=1}^n\left(h(x_i)-y_i\right)^2=\eta$$

3　基于函数集信息量的模型选择

3.1 不含噪声样本的模型选择问题

上节表明，函数集信息量 $Q(\Phi,Z)$ 从量上直观地反映了 $g(x)$ 相对于样本集 Z 在函数集 Φ 上的最佳逼近程度，定理还告诉了模型选择的基本思想：根据给定的函数集 Φ 和样本集 Z，从 Φ 中选择合适的预测模型函数子集 $\Phi(\vartheta)=\{\phi(x,\theta_i),\theta_i\in\Theta\}_{i=1}^q$，尽可能最大化函数集信息量 $Q(\Phi(\vartheta),Z)$，其中，$\vartheta=\{\theta_i\}_{i=1}^q$ 是与模型选择相关的参数集。设预测模型：

$$f(x,\vartheta)=\sum_{i=1}^q\alpha_i\phi(x,\theta_i)\tag{1}$$

记 $\alpha(\vartheta)=[\alpha_1,\cdots,\alpha_n]^{\mathrm{T}}$ 为预测模型系数向量，则

$$\alpha(\vartheta)=\left(B^{\mathrm{T}}(\vartheta)B(\vartheta)\right)^{-1}B^{\mathrm{T}}(\vartheta)y\tag{2}$$

其中 $B(\vartheta)=\left[\varphi(\theta_1),\cdots,\varphi(\theta_q)\right]$ 为列满秩阵。

实际上，函数集 Φ 是影响 $Q(\Phi(\vartheta),Z)$ 的关键因素之一，因此，选择合适的函数集 Φ 也是模型选择的关键；然而，至今国内外的最新研究在理论上尚没有突破如何根据给定的样本集 Z 确定合适的函数集 Φ 的难题，在此也仅讨论已知函数集 Φ 时，如何选择合适的函数子集 $\Phi(\vartheta)$ 的问题。

3.2　含噪声样本的模型选择问题

然而，在系统辨识和信号处理等领域中，往往不能直接得到系统的真实输出，只能得到系统含噪声的观测输出。下面主要研究一类基于含噪声样本的模型选择问题。

定理 4　设系统观测输出满足 $y=g(x)+\varepsilon$，其中，$g(x)$ 为未知的系统真实输出，ε 为服从独立同分布的零均值白噪声，且已知噪声方差 $\mathrm{var}(\varepsilon)$；记 $Z^*=\{(x_i,g(x_i))\}_{i=1}^n$，$Z_\varepsilon=\{(x_i,\varepsilon_i)\}_{i=1}^n$，则有

$$\left.\begin{aligned}\lim_{n\to\infty}Q(\Phi(\vartheta),Z)&=\lim_{n\to\infty}\left\{\ln\frac{A_2}{\mathrm{var}(\varepsilon)}-\ln\left[\left(\frac{A_2}{\mathrm{var}(\varepsilon)}-1\right)\right.\right.\\&\left.\left.\cdot\exp\left(-Q(\Phi(\vartheta),Z^*)\right)+\exp\left(-Q(\Phi(\vartheta),Z_\varepsilon)\right)\right]\right\}\\\lim_{n\to\infty}Q(\Phi(\vartheta),Z_\varepsilon)&=\lim_{n\to\infty}\left\{-\ln\left(1-\frac{q}{n}\right)\right\}\end{aligned}\right.\tag{3}$$

成立。

推论 1　$\max\lim\limits_{n\to\infty}Q(\Phi(\vartheta),Z)$ 等价于 $\max\lim\limits_{n\to\infty}Q(\Phi(\vartheta),$

Z^*) 且 $\min\lim_{n\to\infty} Q\left(\Phi(\vartheta), Z_\varepsilon\right)$ 的充要条件： $\lim_{n\to\infty}\frac{q}{n}=0$ 。

推论 2 设 $\Phi(\vartheta)$ 满足 $\lim_{n\to\infty}\frac{q}{n}=0$，若 $\lim_{n\to\infty} Q\left(\Phi(\vartheta), Z\right)=\lim_{n\to\infty}\ln\frac{A_2}{\mathrm{var}(\varepsilon)}$，则 $\int\left[f(x,\vartheta)-g(x)\right]^2\mathrm{d}x=0$。然而，上述分析只是在样本数为无穷大时才有意义。对于有限样本近似认为定理和推论成立，因此，有限样本下模型选择问题可以近似为在

$$Q\left(\Phi(\vartheta), Z\right)\le \ln\frac{A_2}{\mathrm{var}(\varepsilon)}-\ln\left(1-\frac{q}{n}\right) \tag{4}$$

约束下优化如下问题：

$$\hat{\vartheta}\approx \arg\max_{\vartheta} Q\left(\Phi(\vartheta), Z\right)\cup\min_{\vartheta} q \tag{5}$$

其中，式（4）对 $Q\left(\Phi(\vartheta), Z\right)$ 的约束是有意义的，它不仅能够有效地抑制由于样本集 Z 的有限性和随机性可能导致模型选择的早熟现象，同时也能有效地抑制过学习问题。然而，实际上最大化 $Q\left(\Phi(\vartheta), Z\right)$ 和最小化 q 往往相互矛盾，因此有限含噪声样本的模型选择问题可以归结为：在式（4）的约束下，尽可能从 Φ 中选择少量的元素组成 $\Phi(\vartheta)$ 并最大化 $Q\left(\Phi(\vartheta), Z\right)$。

有限含噪声样本的模型选择过程将不可避免地把噪声信息引入预测模型中，在大样本情况下，若 $\Phi(\vartheta)$ 满足 $q<<n$ 时，预测模型中包含的噪声信息微乎其微，且式（4）可以近似简化为 $Q\left(\Phi(\vartheta), Z\right)\le \ln\frac{A_2}{\mathrm{var}(\varepsilon)}$；然而对于小样本来说，由于样本数 n 较小，往往很难保证 $\Phi(\vartheta)$ 满足 $q<<n$，因此，预测模型中包含较多的噪声信息。

3.3 一种次优的模型选择算法

由问题可知，从 Φ 中选择 $\Phi(\vartheta)$ 等价于从 Θ 中选择 ϑ，考虑到 Θ 的元素往往很多，因此从中寻找最优 ϑ 将会耗费大量计算时间，这里给出一种简单的次优迭代算法：

(1) 初始化 $\vartheta=\varnothing$，设置 $Q\left(\Phi(\vartheta), Z\right)$ 的下限 Q_0，以及 q 的上限 q_0；

(2) 从 Θ 中选择某一向量 $\theta\notin\vartheta$ 加入 ϑ 中，在满足式(4)约束条件下，使得 $\max_{\vartheta=\vartheta+\theta} Q\left(\Phi(\vartheta), Z\right)$ 成立；

(3) 若 Θ 中任意 $\theta\notin\vartheta$ 加入 ϑ 中，下列任一情况出现时 (a) 不再满足式（4）的约束条件；(b) $Q\left(\Phi(\vartheta), Z\right)$ 增加量很小；(c) $\left\{\varphi(\theta),\theta\in\vartheta\right\}$ 各向量线性相关；(d) $q\ge q_0$；

转到第(4)步，否则转到第(2)步；

(4) 从 ϑ 中删除某一基向量 $\theta\in\vartheta$，且使得 $\max_{\vartheta=\vartheta-\theta} Q\left(\Phi(\vartheta), Z\right)$ 成立；

(5) 若 $Q\left(\Phi(\vartheta), Z\right)>Q_0$，转到第(4)步开始；否则仍保留上一步删除的向量，算法结束。

由函数集信息量定义可知，由于 $\left\{\varphi(\theta),\theta\in\vartheta\right\}$ 中各元素并不一定正交，从对信息量 $Q\left(\Phi(\vartheta), Z\right)$ 的贡献来说，算法第(2)步所选择的 ϑ 中某些元素的贡献相对小些或它所包含的信息大部分已包含于其它元素中，因此算法的第(4)步非常巧妙，也是必要的，它可以删除那些对信息量贡献不大的冗余元素，从而提高预测模型的泛化性能。

若 Θ 为有限个元素的集合，则上述算法在有限次迭代后结束。若 Θ 为连续空间域，设算法第(2)步每次迭代时，选择某个未知向量 $\theta\notin\vartheta$ 加入 ϑ 中，记 $Q(\theta)=Q\left(\Phi(\vartheta), Z\right)$，则 $\partial Q(\theta)/\partial\theta=0$，且 $\partial^2 Q(\theta)/\partial\theta^2\le 0$；然而，对于实际问题来说，求解上述偏导有时显得非常困难，因此，可以把 Θ 的连续空间域离散化为有限个元素的集合，从而把复杂的问题简单化。

4 数值试验

下面通过数值试验验证上述模型选择理论和方法的可行性和有效性。设对象 $y=\exp\left(-x^2/20\right)\cdot\cos x+\varepsilon$，$x\in[-4\pi,4\pi]$，已知 $\mathrm{var}(\varepsilon)=0.001$；在其中均匀取 $n=252$ 个样本。

设 $\phi(x,\theta)=\exp\left(-(x-\theta)^2/\sigma^2\right)$，其中，$\sigma^2=9.0$，$\Theta$ 为区间 $[-4\pi,4\pi]$。为了简单起见，离散化连续参数域 Θ，在其中均匀取 $l=252$ 个参数样本组成新的有限参数集 $\Omega=\left\{\theta_i\right\}_{i=1}^{l}$，再利用上节算法，结果选取的 $E(\vartheta)$ 中包含 8 个元素，其中，$\vartheta=\{0, 2.8923, -1.9947, -3.8896, 4.7872, -5.8843, 2.7925, -7.2805\}$，预测模型输出与实际样本对比如图 1 所示。

从图 1 可以看出，上述算法建立的预测模型取得了非常满意的效果，有很强的抗噪声能力，完全克服了传统的模型选择算法可能存在的欠学习或过学习以及局部最小点问题，同时算法在迭代过程中具有很好的一致收敛性。但多次重复试验表明，噪声和样本分布直接影响选择的预测模型函数子集，尤其当上述算法中指标要求较高的情况下，后续选择的次要元素差别较大，因此，实际应用中适当放宽指标要求，可以有效地提高预测模型抗噪声性能。

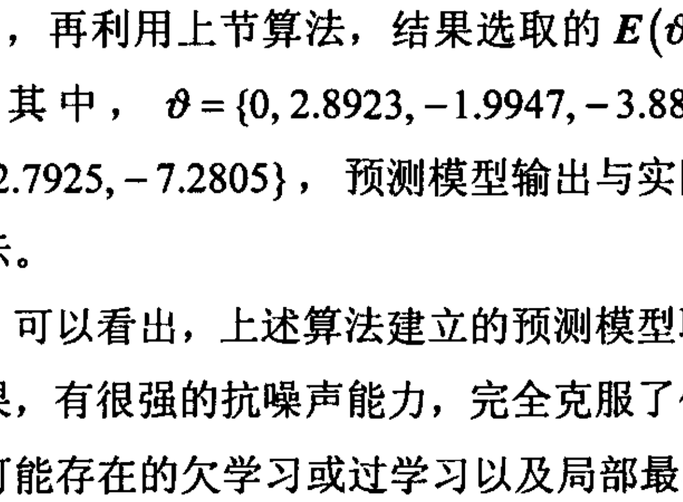

图 1 预测模型输出与实际样本比较

5　结论

本文针对模型选择问题，首先提出了子空间信息量概念，在此基础上引出了函数集信息量概念，详细讨论了基于函数集信息量概念的模型选择的相关理论和方法，利用它能很好地解决模型选择领域中存在的诸多难题。这种理论和方法完全符合统计学习理论所提出的结构风险最小化原则，具有很强的抗噪声和小样本学习能力，模型也具有很好的推广性，克服了传统的模型选择方法存在的诸多弊端，因此，它具有较好的理论和应用价值。同时，它可以广泛地推广应用到其它机器学习领域中。

参 考 文 献

[1] 张学工. 关于统计学习理论与支持向量机. 自动化学报, 2000, 26(1): 32 – 42.

[2] Sugiyama M, Ogawa H. Subspace information criterion for model selection. *Neural Computation*, 2001, 13(8): 1863 – 1889.

[3] Stolke A. Bayesian learning of probabilistic language models. [Ph.D. Dissertation], University of California, Berkeley, 1994.

[4] Hemant Ishwaran, Lancelot F, Jiayang Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 2001, 96(456): 1316 – 1332.

[5] Cherkassky V, Shao X, Mulier F M, Vapnik V N. Model complexity control for regression using VC generalization bounds. *IEEE Trans. on Neural Networks*, 1999, 10(5): 1075 – 1089.

[6] Barron A R, Cover T M. Minimum complexity density estimation. *IEEE Trans. on Information Theory*, 1991, 37(4): 1034 – 1054.

[7] Yamanishi K. A decision-theoretic extension of stochastic complexity and its application to learning. *IEEE Trans. on Information Theory*, 1998, 44(4): 1424 – 1439.

[8] Wood S N. Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. Royel Statist. Soc. B*, 2000, 62(1): 413 – 428.

[9] Chapelle O, Vapnik V N, Bengio Y. Model selection for small-sample regression. *Machine Learning Journal*, 2002, 48(1): 9 – 23.

[10] Hurvich C M, Tsai C L. Regression and time series model selection in small samples. *Biometrika*, 1989, 76(13): 297 – 307.

[11] Bousquet O, Elisseeff A. Stability and generalization. Journal of Machine Learning Research 2, 2002: 499 – 526.

[12] Konishi S, Kitagawa G. Generalized information criterion in model selection. *Biometrika*, 1996, 83(4): 875 – 890.

[13] Akaike H. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 1974, AC-19(6): 716 – 723.

[14] Hurvich C M, Tsai C L. Bias of the corrected AIC criterion for under-fitted regression and time series models. *Biometrika*, 1991, 78(2): 499 – 509.

[15] Murata N, Yoshizawa S, Amari S. Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Trans. on Neural Networks*, 1994, 5(6): 865 – 872.

[16] Shibata R. Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica*, 1997, 7(2): 375 – 394.

盛守照：　男，1977年生，博士生，研究方向：机器学习、智能控制、先进无人机飞行控制等.

王道波：　男，1957年生，教授，博士生导师，研究方向：先进无人机飞行控制、机电模拟等.