第 31 卷第 5 期
2014 年 5 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 31 No. 5
May 2014

DOI: 10.7641/CTA.2014.20428

基于Vapnik-Chervonenkis泛化界的极限学习机模型复杂性控制

刘学艺[1,2], 宋春跃[3†], 李　平[1,3]

(1. 浙江大学 航空航天学院, 浙江 杭州 310027; 2. 中国计量学院 数学系, 浙江 杭州 310018;
3. 浙江大学 工业控制研究所, 浙江 杭州 310027)

摘要: 模型复杂性是决定学习机器泛化性能的关键因素, 对其进行合理的控制是模型选择的重要原则. 极限学习机(extreme learning machine, ELM)作为一种新的机器学习算法, 表现出了优越的学习性能. 但对于如何在ELM的模型选择过程中合理地度量和控制其模型复杂性这一基本问题, 目前尚欠缺系统的研究. 本文讨论了基于Vapnik-Chervonenkis(VC)泛化界的ELM模型复杂性控制方法(记作VM), 并与其他4种经典模型选择方法进行了系统的比较研究. 在人工和实际数据集上的实验表明, 与其他4种经典方法相比, VM具有更优的模型选择性能: 能选出同时具有最低模型复杂性和最低(或近似最低)实际预测风险的ELM模型. 此外, 本文也为VC维理论的实际应用价值研究提供了一个新的例证.

关键词: VC泛化界; 模型复杂性; 极限学习机; 小样本; 实际预测风险

中图分类号: TP18　　**文献标识码**: A

Model complexity control of extreme learning machine using Vapnik-Chervonenkis generalization bounds

LIU Xue-yi[1,2], SONG Chun-yue[3†], LI Ping[1,3]

(1. School of Aeronautics and Astronautics, Zhejiang University, Hangzhou Zhejiang 310027, China;
2. Department of Mathematics, China Jiliang University, Hangzhou Zhejiang 310018, China;
3. Institute of Industrial Process Control, Zhejiang University, Hangzhou Zhejiang 310027, China)

Abstract: Model complexity is the critical element in determining the generalization ability of a learning machine. Hence, any learning machine needs to have proper provisions for complexity control. Extreme learning machine (ELM) has recently become increasingly popular due to its high learning speed and good generalization performance. However, there is a lack of systematic study on how to accurately measure and control its complexity for the purpose of good generalization. In this paper, the Vapnik-Chervonenkis (VC) bound-based model selection method (VM) is discussed, and then is compared with other 4 classic statistical model selection criteria. Simulations on the artificial and real-world datasets show that VM can achieve the best model selection performance among all 5 model selection methods; it provides the optimal ELM model with both the lowest model complexity and the smallest or nearly smallest real prediction risk. In addition, this paper also provides a strong evidence for the practical applicability of VC generalization bound in terms of model selection.

Key words: VC generalization bounds; model complexity; extreme learning machine; small sample; real prediction risk

1　引言(Introduction)

前向神经网络, 如多层感知器(multilayer perceptron)和径向基函数(radial basis function, RBF)网络, 是神经网络研究领域中最经典的网络模型, 其中, 单隐层前馈神经网络(single-hidden-layer feed-forward net-work, SLFN)由于结构简单易实现, 同时具有强大的非线性逼近能力, 已得到深入的研究和广泛的应用[1]. 极限学习机(extreme learning machine, ELM)是近几年提出的一种针对单隐含层前向神经网络的机器学习算法[2–3]. ELM算法中首先随机选取SLFN的输入层权重, 然后利用最小二乘思想计算输出层权重. 模型训练时间显著优于基于梯度信息的传统迭代算法以及支持向量机(support vector machine, SVM)算法, 同时却能够获得可比拟SVM的泛化性能[3–4]. 由于其训练速度快、易于实现以及泛化能力强等鲜明的特点, 正吸引着越来越多学者的关注[5–9].

统计学习理论表明, 学习机器的泛化能力取决于学习机器所实现的函数集的容量(即模型复杂性). 因

收稿日期: 2012−04−28; 录用日期: 2014−02−26.
†通信作者. E-mail: cysong@iipc.zju.edu.cn; Tel.: +86 571-87951442.
基金项目: 国家重点基础研究发展计划“973”资助项目(2009CB320603); 国家自然科学基金资助项目(61273085); 浙江省自然科学基金资助项目(LY14F030020).

此, 在模型参数选择的过程中合理地控制模型复杂性, 对其泛化性能的提高至关重要. 然而, 对于ELM, 如何在参数选择过程中合理地度量和有效地控制其模型复杂性, 目前尚欠缺系统的研究. 不同的模型选择标准会得到不同的模型复杂性控制效果, 并将导致算法性能上的差异, 因此, 研究适用于ELM算法的模型选择方法具有重要意义. 当前常用的ELM模型参数选择方法有交叉验证方法[4–7]、统计量和信息增益[8]、Akaike信息准则(Akaike information criterion, AIC)[9]等. 其中, 交叉验证方法应用最多, 如k重交叉验证和留一交叉验证等. 除以上方法外, Bayesian信息准则(Bayesian information criterion, BIC)[10]、Shibata模型选择标准(Shibata's model selector, SMS)[11]等, 作为基于统计理论提出的经典模型选择方法同样可以应用于ELM.

另一方面, VC 理论(Vapnik-Chervonenkis theory)是专门针对小样本情形下的机器学习问题所提出的[12]. Vapnik据此进一步提出的结构风险最小化原则, 为模型复杂性控制提供了更为一般且有效的理论框架. VC泛化界(VC generalization bound)是其中最核心的概念, 它给出了模型实际泛化性能上界的一种分析估计方法. 基于VC泛化界的模型选择方法已经吸引了许多学者的研究兴趣, 比如: Cherkassky等[13]、Cherkassky等[14]和Shao等[15]通过仿真实验发现, 该方法优于其他模型选择方法; Cherkassky等[16]成功地将其应用于小波信号降噪问题; Chapelle等[17]提出一种基于VC理论的小样本问题模型选择方法, 并通过仿真实验验证了所提方法的优越性.

然而, 由于对VC泛化界本身保守性的顾虑[18]以及VC维计算的困难性[18–19]等原因, 目前它仅在有限且简单的几个算法中获得应用. 因此, 相比于VC泛化界的重要理论意义, 其实际应用价值仍是当前需要进一步研究的问题. 注意到, 对于ELM, 文献[20]已经对其VC维大小进行了严格证明. 本文将在此基础上引入基于VC泛化界的ELM模型复杂性控制方法, 并通过仿真实验比较该方法与其他4种经典ELM模型选择方法的性能差异. 这一方面能为ELM的模型选择提供指导, 同时也为研究VC 泛化界的实际应用价值提供有益线索. 为简单起见, 本文以回归为例讨论, 得到的结果可推广到分类情形.

2　极限学习机(ELMs)

2.1　学习问题的一般提法(The learning problem)

假设给定有限样本$(\boldsymbol{x}_i, y_i)(i=1,\cdots,N)$, 估计如下未知函数:

$$y = f(\boldsymbol{x}) + \varepsilon, \tag{1}$$

其中ε为零均值随机误差. 一个学习算法总是试图利用训练样本从逼近函数集合$\{f(\boldsymbol{x},\theta)|\theta\in\Theta\}$中发现"最好"的模型$f(\boldsymbol{x},\theta_0)$, 以最小化如下实际预测风险泛函:

$$R(\theta) = \int L(y, f(\boldsymbol{x},\theta))p(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y, \tag{2}$$

其中, $p(\boldsymbol{x},y)$是样本的联合概率密度函数; $L(y,f(\boldsymbol{x},\theta))$是预测损失函数. 预测风险泛函衡量学习算法对未知目标函数$f(\boldsymbol{x})$的预测精度, 对回归问题常采用二次损失形式, 即$L(y,f(\boldsymbol{x},\theta))=(y-f(\boldsymbol{x},\theta))^2$, 此时, 实际预测风险泛函变为

$$R(\theta) = \int (y-f(\boldsymbol{x},\theta))^2 p(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y. \tag{3}$$

经验风险最小化方法直接最小化经验风险, 即在训练样本点上的平均损失

$$R_{\mathrm{emp}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i,\theta))^2. \tag{4}$$

2.2　极限学习机(ELMs)

极限学习机采用如下函数集来估计式(1)中的未知函数:

$$\{f(\boldsymbol{x}) = \sum_{j=1}^{L}\beta_j G(\boldsymbol{a}_j, b_j, \boldsymbol{x}) | \boldsymbol{a}_j\in\mathbb{R}^n;\ \beta_j, b_j\in\mathbb{R}\}, \tag{5}$$

其中: L为单隐层前向神经网络的隐含层神经元个数; $\boldsymbol{a}_j=(a_{1j},a_{2j},\cdots,a_{nj})^{\mathrm{T}}$为连接第$j$个隐含层结点的输入权值向量; b_j为第j个隐层神经元的阈值; β_j为连接第j个隐层结点的输出权值; 且

$$G(\boldsymbol{a}_j,b_j,\boldsymbol{x}) = \begin{cases} g(\boldsymbol{a}_j^{\mathrm{T}}\boldsymbol{x}+b_j), & g(\,)\text{为加性函数,} \\ g(\dfrac{\|\boldsymbol{x}-\boldsymbol{a}_j\|^2}{b_j}), & g(\,)\text{为RBF函数.} \end{cases} \tag{6}$$

$g(\,)$为隐层神经元激活函数.

对于给定的N个样本$\{(\boldsymbol{x}_i,y_i)\}_{i=1}^N$, ELM算法首先随机取定$\boldsymbol{a}_j$和$b_j$, 然后通过最小化经验风险式(4)来估计$\boldsymbol{\beta}=[\beta_1\ \beta_2\ \cdots\ \beta_L]^{\mathrm{T}}$. 注意到式(5)关于未知参数呈线性关系, 故可转而求解如下线性方程组:

$$\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{y}, \tag{7}$$

其中, 隐含层输出矩阵

$$\boldsymbol{H} = \begin{bmatrix} G(\boldsymbol{a}_1,b_1,\boldsymbol{x}_1) & \cdots & G(\boldsymbol{a}_L,b_1,\boldsymbol{x}_1) \\ \vdots & & \vdots \\ G(\boldsymbol{a}_1,b_N,\boldsymbol{x}_N) & \cdots & G(\boldsymbol{a}_L,b_N,\boldsymbol{x}_N) \end{bmatrix}_{N\times L}, \tag{8}$$

且$\boldsymbol{y}=[y_1\ \cdots\ y_N]^{\mathrm{T}}$.

对于隐含层输出矩阵$\boldsymbol{H}$, 若$L\leqslant N$, 则$\boldsymbol{H}$以概率1列满秩[3]. 同时, Huang等[3]指出: 对绝大多数问题, 都有$L\leqslant N$. 至此, 输出层权重$\boldsymbol{\beta}$可以由式(7)的极小范数解进行估计, 即

$$\hat{\boldsymbol{\beta}} = \boldsymbol{H}^{\dagger}\boldsymbol{y}, \tag{9}$$

式中$\boldsymbol{H}^{\dagger}$为$\boldsymbol{H}$的Moore-Penrose广义逆.

2.3 极限学习机的VC维(VC dimension of ELMs)

VC维的定义可叙述为[12]: 对一个指示函数(只有0和1两种取值的函数)集, 如果存在N个样本能够被函数集中的函数按所有可能的2^N种形式分开, 则称函数集能把这N个样本打散; 函数集的VC维就是它能打散的最大样本数目N. VC维是统计学习理论的核心概念, 对于确定学习机器的泛化能力、训练集规模等具有重要作用. 遗憾的是, 目前还没有关于任意函数集VC维计算的通用理论, 只能得到一些较特殊函数集的VC维. 例如, N维实空间中的线性分类器和线性实函数集的VC维是$N+1$; 函数集$\{\sin(ax)|a\in\mathbb{R}\}$的VC维是无穷大.

对于神经网络(特别是前馈神经网络)的VC维, 目前已取得了一定的研究成果[21–24]. 但是这些结果还是以范围较宽的上下界估计为主, 不具备实用性.

对于ELM, 虽然它是一种针对SLFN的学习算法, 但是目前关于神经网络VC维的已有结论并不能直接应用于ELM算法. 原因在于VC维的大小会同时受到神经网络的结构、激活函数和学习算法的影响, 而ELM不仅在算法上显著不同于基于梯度下降的传统SLFN算法, 而且它的激活函数可选范围也更为广泛[25–27]. 为此, 文献[20]对ELM的VC维进行了讨论, 并证明: ELM模型的VC维以概率1等于其隐含层神经元个数. 这为基于VC维对ELM的泛化性能进行评估并进行模型选择提供了可能.

3 极限学习机的模型复杂性控制(Model complexity control of ELMs)

学习问题的目的是获得一个对未来样本具有最小预测风险的模型. 对于一组给定的训练样本, 一定存在一个具有最优模型复杂性的模型具有最小的预测风险. 因而, 设计一个学习算法使其能通过合理地模型选择策略获得最佳模型, 本质上就是模型复杂性的控制问题. 而要控制模型复杂性, 一个根本性的问题是如何定义式(3)对应的实际预测风险. 认识到这一点, 下面分别给出本文要讨论的几种模型选择方法对式(3)的定义方式.

3.1 经典模型选择方法(Classical model selection criteria)

经典的模型选择方法都是基于线性模型的渐近结果提出, 可分为两类: 数据重采样方法和分析估计方法. 前者主要指交叉验证方法(本文考虑其中的留一交叉验证算法, 记作LOO). 而后者则根据模型复杂性对经验风险进行必要的惩罚来估计实际预测风险, 它们都可统一为如下形式:

$$R_{\rm est}(d)=R_{\rm emp}(d)\cdot r(d,N), \tag{10}$$

或者

$$R_{\rm est}(d)=R_{\rm emp}(d)+r(d/N,\sigma^2), \tag{11}$$

其中r常被称作惩罚因子, 关于d/N单调递增[28]. 本文考虑如下3种分析估计方法:

1) Akaike信息准则(AIC).

$$AIC(d)=R_{\rm emp}(d)+\frac{2d}{N}\hat{\sigma}^2; \tag{12}$$

2) Bayesian信息准则(BIC).

$$BIC(d)=R_{\rm emp}(d)+(\ln N)\frac{d}{N}\hat{\sigma}^2; \tag{13}$$

3) Shibata模型选择准则(SMS).

$$SMS(d)=R_{\rm emp}(d)+2\frac{d}{N}R_{\rm emp}(d). \tag{14}$$

在以上3种标准中, d表示模型中自由参数的个数(对于ELM, d可取为隐含层神经元个数L); σ是式(1)中噪声的标准偏差. AIC, BIC和SMS3种预测风险估计都建立在渐进分析理论的基础上, 对于足够大的样本数量N, 它们是渐进等价的. 在实际应用中, 噪声标准偏差σ是未知的, 本文利用下式对其进行估计:

$$\hat{\sigma}^2=\frac{N}{N-d}\cdot\frac{1}{N}\sum_{i=1}^{N}\left(y_i-f(\boldsymbol{x}_i,\theta)\right)^2. \tag{15}$$

将式(15)代入式(12), 可得

$$AIC(d)=\frac{1+d/N}{1-d/N}R_{\rm emp}(d). \tag{16}$$

此式即为最终预测误差(final prediction error, FPE)[29]模型选择标准.

显然, 取定$d=L$后, 即可由式(13)–(14)(16)估计ELM的实际预测误差, 并进行模型选择.

需要指出, 以上3种模型选择标准都需要以下基本假设: 1) 目标函数线性; 2) 容许函数集中包含目标函数, 即学习机器能得到目标函数的无偏估计; 3) 噪声独立同分布; 4) 能够获得全局最小化的经验风险. 而在绝大多数实际应用中, 这些假设并不成立.

3.2 基于VC理论的模型复杂性控制[20](Model complexity control using VC generalization bounds)

根据统计学习理论, 结构风险最小化方法需要构造所有容许函数集合的一个嵌套结构: $S_1\subset S_2\subset\cdots\subset S_k\subset\cdots$, 它们对应的VC维单调递增, 即: $h_1\leqslant h_2\leqslant\cdots\leqslant h_k\leqslant\cdots$.

对于ELM, 容许函数集由式(5)给定. 若记$f_L(\boldsymbol{x})=\sum_{j=1}^{L}\beta_jG(\boldsymbol{a}_j,b_j,\boldsymbol{x})$, 则$\{f_L(\boldsymbol{x})\}$满足

$$\{f_1(\boldsymbol{x})\}\subset\{f_2(\boldsymbol{x})\}\subset\cdots\subset\{f_L(\boldsymbol{x})\}\subset\cdots. \tag{17}$$

显然, 函数集$\{f_L(\boldsymbol{x})\}$的VC维为L, 关于L单调递增. 因而, 式(17)就给定了一个满足条件的嵌套结构.

对于给定的训练样本, 结构风险最小化原则通过以下步骤估计具有最优复杂性的模型[12]:

1) 对每个子集S_k最小化经验风险式(4).

2) 对每个子集S_k估计实际预测风险式(3).

3) 通过最小化实际预测风险估计最优模型.

其中, 实际预测风险通常可由交叉验证、AIC准则等方法来估计, 本文将重点研究基于VC泛化界的估计方法. VC泛化界是经验风险$R_{\rm emp}(\theta)$、训练样本数量N和VC维h的函数. 对于回归情形, 下面的VC泛化界不等式给出了实际预测风险的上界估计, 且以概率$1-\eta$成立:

$$R(h)\leqslant R_{\rm emp}(h)(1-c\sqrt{\frac{h(\ln(\frac{aN}{h})+1)-\ln\eta}{N}})_+^{-1}. \tag{18}$$

在此, 首先要确定(18)中的常数a, c以及置信水平$1-\eta$. 根据Vapnik等[12]和Cherkassky等[14] 的相关理论分析和实验验证, 一种合理的取值策略是: $a=1$, $c=1$, $\eta=1/\sqrt{N}$. 由此, 式(18)可简化为

$$R(h)\leqslant R_{\rm emp}(h)(1-\sqrt{p-p\ln p+\frac{\ln N}{2N}})_+^{-1}, \tag{19}$$

其中: $p=h/N$, h为VC维.

针对某些学习机器(特别是线性学习算法), VC泛化界的模型选择效果已经得到实验验证. ELM作为一种新的SLFN算法, 本质上也可视作一种特殊的线性算法: 经由ELM隐含层确定的随机特征映射, 将原始样本$\boldsymbol{x}$投影到某个新的特征空间, 然后在该特征空间中基于线性算法完成求解. 对于这种建立在随机特征空间上的特殊线性算法, 能否基于VC维对其模型复杂性进行有效的度量, 并进一步利用VC泛化界进行模型选择, 是目前值得深入研究的问题.

4　仿真实验(Simulations)

本文在仿真实验部分共考虑了5种ELM模型选择标准: 基于VC泛化界的模型选择标准(记作VM)、3种基于统计方法的分析估计方法(AIC, BIC, SMS)和一种样本重采样方法(留一交叉验证, 记作LOO). 为了更客观地评价以上方法的性能, 本文将分别针对人工数据和实际数据进行仿真实验. 为了讨论方便, 在仿真实验中将采取网格搜索的策略对未知参数(隐含层神经元个数L)进行遍历, 这种方法虽然效率不高但并不影响对5种模型选择标准的性能评价(在实际应用中, 可以基于现代优化方法设计更加快速的模型选择方法).

4.1　人工数据(Artificial cases)

针对人工数据的仿真实验, 具体设计如下: 首先, 生成训练样本数据. 对给定的N和正态分布噪声标准差σ, 生成100个训练样本$\{(x_i,y_i)\}_{i=1}^N$, 其中, 自变量x在区间$[0,1]$上随机产生, 对应函数值为$y=h(x)+\varepsilon$, $\varepsilon\sim N(0,\sigma^2)$. 这里$h(x)$取如下两个目标函数:

Sine-squared(SS)函数:

$$h_1(x)=\sin^2(2\pi x),\ x\in[0,1]. \tag{20}$$

分段多项式(piecewise polynomial, PP)函数:

$$h_2(x)=\begin{cases}4x^2(3-4x), & x\in[0,\frac{1}{2}],\\ \frac{4}{3}x(4x^2-10x+7)-\frac{3}{2}, & x\in(\frac{1}{2},\frac{3}{4}],\\ \frac{16}{3}x(x-1)^2, & x\in(\frac{3}{4},1].\end{cases} \tag{21}$$

以上两个函数(图1给出了它们的图形)经常被用来评价模型选择方法的优劣[14, 17, 30].

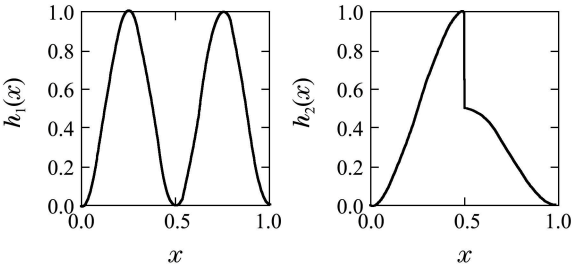


图 1　仿真实验中的两个目标函数

Fig. 1　Two target functions used in simulations

然后, 分别由5种不同标准选择“最优”模型: 用隐含层结点数为$L(L=1,2,\cdots,25)$的ELM模型分别对一组训练样本进行学习, 并由5种模型选择标准(AIC, BIC, SMS, LOO和VM)选出5个“最优”模型(即最优的L值); 并对每一组样本重复此工作, 可得到5组(每组100个)最优L值.

最后, 估计并统计每组最优模型的实际预测风险: 按训练样本的生成方式生成1000个测试样本, 再由此计算并统计每组100个“最优”模型的预测均方误差(记作MSE).

为了综合评价VM及其他模型选择方法的性能, 仿真实验综合考虑了如下3种因素: 1) 训练样本数量: 小样本(30)、中等规模样本(10^2)和大样本(10^3, 10^4)3种情况; 2) 信噪比(signal-to-noise ratio, SNR): 分别取0.5, 2和4这3种水平; 3) 激活函数类型: Sigmoid和RBF两种.

以上实验结果以盒须图形式进行统计分析, 其中, 矩形中3条水平线段由下而上分别表示下侧四分位数、中位数和上侧四分位数; 由矩形延伸出去的“须”长度为1.5倍四分位距; 为了结果显示和比较方便起见, 人工数据集情形下的异常值未在图中标出.

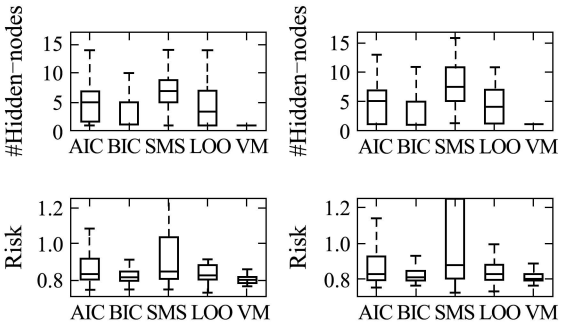
1) 小样本情形.

针对小样本($N=30$)情形, 本文对如上所述的所有情形进行了12组仿真实验, 详细比较了VM方法与其他方法在性能上的差异.

考虑到隐含层神经元个数L和实际预测风险是极

限学习机模型选择过程中最重要的两个因素, 图2–7给出了12组实验中这两个因素(分别用# Hidden-nodes和Risk表示)的盒须图. 从这些图形中容易得到如下结论: ① 当训练样本个数为30时, VM获得了更低的实际预测风险(在6个图形对应的12组实验中, VM都得到了更低的$\frac{3}{4}$分位点和上须点). 也就是说, 对于小样本情形, VM显著优于其他模型选择方法; ② VM得到的最优ELM模型具有最少的隐含层神经元, 这说明VC泛化界在模型选择时能够合理地考虑模型复杂性增加带来的置信风险, 选出具有最优泛化性能的模型; ③ VM得到的实际预测风险盒须图的矩形高度最小(实际上, 包括上下须长部分对应的高度同样最小), 这说明利用VC泛化界来估计实际预测风险具有明显的优势; ④ 在3种分析估计方法中, BIC标准也倾向于选择具有更少隐含层神经元的ELM模型, 并表现出优于AIC和SMS的性能, 但性能明显比VM差; ⑤ LOO作为一种重采样方法, 能够充分利用有限样本去训练模型, 因此被认为特别适合小样本情形下的模型选择[31]; 以上结果验证了这一点: LOO获得了优于AIC和SMS, 且与BIC相近的性能; ⑥ SMS倾向于选择最多的隐含层神经元个数, 同时得到最差的实际预测风险.

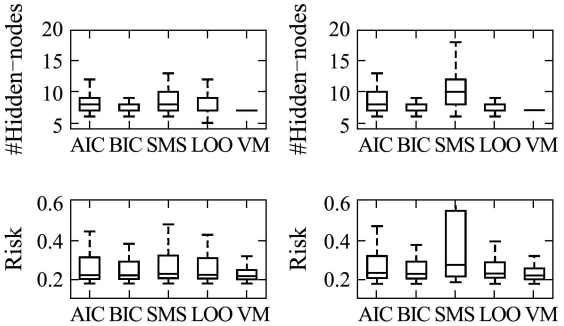
以上仿真结果表明: 对于小样本情形, VM模型选择效果最好, BIC与LOO次之, SMS最差.



(a) Sigmoid (b) RBF

图 2 针对SS函数的仿真结果($N = 30$, SNR $= 0.5$)

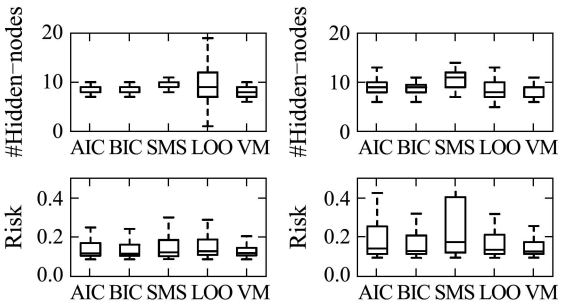
Fig. 2 Results for SS function with $N = 30$ and SNR $= 0.5$



(a) Sigmoid (b) RBF

图 3 针对SS函数的仿真结果($N = 30$, SNR $= 2$)

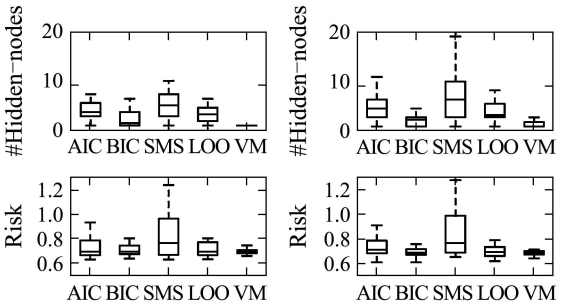
Fig. 3 Results for SS function with $N = 30$ and SNR $= 2$



(a) Sigmoid (b) RBF

图 4 针对SS函数的仿真结果($N = 30$, SNR $= 4$)

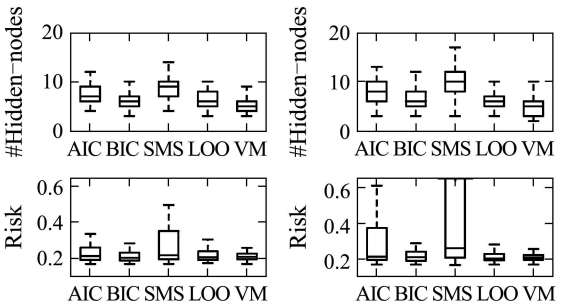
Fig. 4 Results for SS function with $N = 30$ and SNR $= 4$



(a) Sigmoid (b) RBF

图 5 针对PP函数的仿真结果($N = 30$, SNR $= 0.5$)

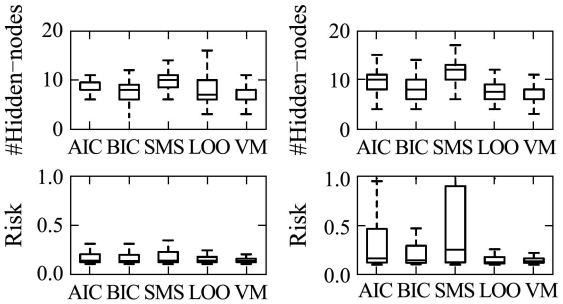
Fig. 5 Results for PP function with $N = 30$ and SNR $= 0.5$



(a) Sigmoid (b) RBF

图 6 针对PP函数的仿真结果($N = 30$, SNR $= 2$)

Fig. 6 Results for PP function with $N = 30$ and SNR $= 2$



(a) Sigmoid (b) RBF

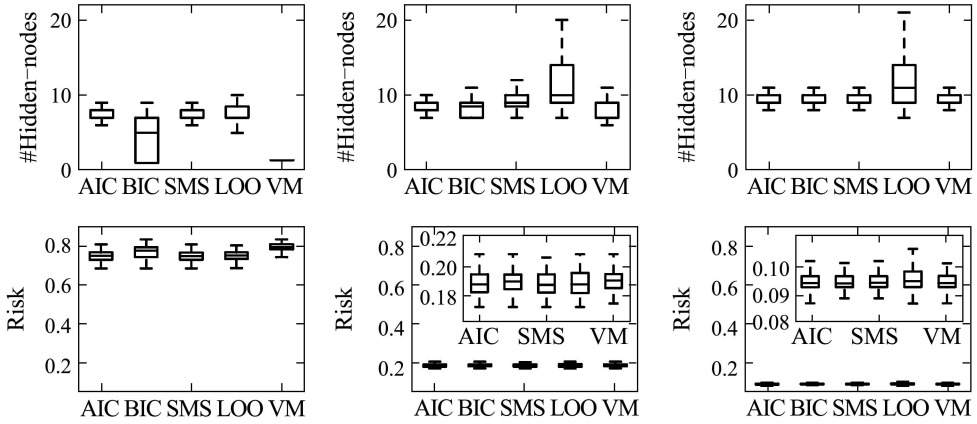
图 7 针对PP函数的仿真结果($N = 30$, SNR $= 4$)

Fig. 7 Results for PP function with $N = 30$ and SNR $= 4$

2) 中等规模样本情形.

针对中等规模样本情形, 本文只考虑Sigmoid激活函数情形. 图8–9给出了当$N = 100$时, 在信噪比SNR分别取0.5, 2和4这3个水平下的比较结果. 图中的比

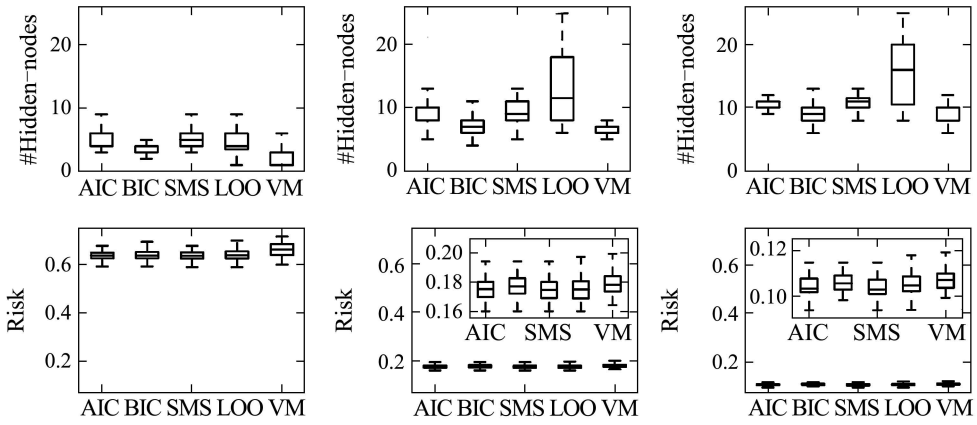
较结果表明: ① 对中等规模样本, VM在选择具有最少的隐含层神经元的ELM模型时, 会在一定程度上增加实际预测风险; 但需要指出, 除SNR = 0.5的情形外, VM的实际预测风险都非常接近于最优预测风险; ② BIC与VM类似, 也侧重于选择低模型复杂性的ELM模型并得到较高的实际预测风险; ③ LOO选择最多的隐含层神经元, 性能变差; ④ 与小样本情形不同, AIC和SMS此时表现出最优的模型选择性能.



(a) Sigmoid, SNR=0.5, N=10^2　(b) Sigmoid, SNR=2, N=10^2　(b) Sigmoid, SNR=2, N=10^2

图 8 针对SS函数的仿真结果(中等规模样本情形)

Fig. 8 Results for SS function under medium sample cases



(a) Sigmoid, SNR=0.5, N=10^2　(b) Sigmoid, SNR=2, N=10^2　(b) Sigmoid, SNR=4, N=10^2

图 9 针对PP函数的仿真结果(中等规模样本情形)

Fig. 9 Results for PP function under medium sample cases

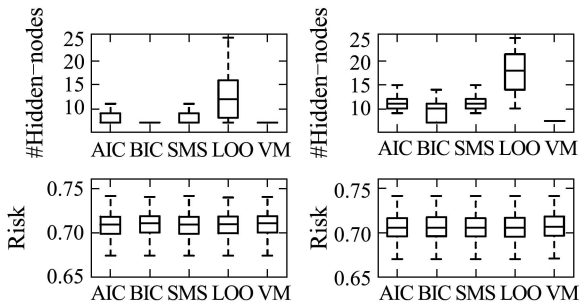
值得说明的是: 在VC泛化界模型选择能力的已有研究[14, 17, 30]中, 对不同的学习算法(如线性算法、小波降噪、k最近邻等), VM在不同的训练样本规模下都表现出了一致的优势. 而以上仿真结果则表明, 当$N = 10^2$时, AIC和SMS的模型选择性能比VC泛化界方法更优越. 其原因可能在于: Huang等[3]指出ELM的泛化性能在最优隐含层神经元个数附近的较大范围内非常稳定(即, 泛化性能对隐含层神经元个数变得不太敏感), 而此时VM会对模型复杂性增加带来的置信风险做出过高的估计.

3) 大样本情形.

针对大样本情形, 本文仅讨论了大噪声的情况(SNR = 0.5). 首先, 图10给出了当$N = 10^3$和10^4时, 对于SS目标函数的比较结果. 结果表明: 当样本数量很大时, ① 5种方法取得了几乎相同的实际预测风险; ② VM和BIC能够选择具有低模型复杂性的ELM模型; ③ 与$N = 10^2$时相似, LOO选择最多的隐含层神经元且最不稳定.

图11给出了另一目标函数的比较结果. 由于该目标函数存在间断点, 客观上需要更多的样本描述其性态. 因此, 观察图11(a)和图11(b)可以发现, 当$N = 10^4$时, 5种模型选择标准才表现出与图10中SS函数情形下相同的规律.

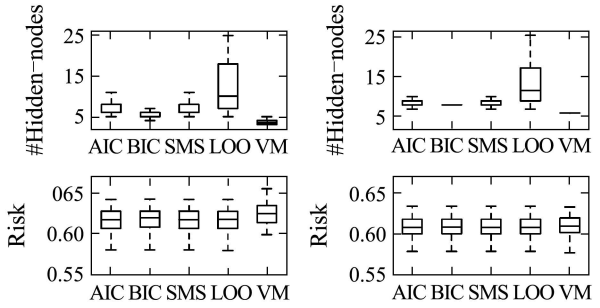
因此, 当训练样本足够大时, 不同方法都会得到相近的实际预测风险, 但在模型复杂性控制方面, VM和BIC方法更具优势.



(a) Sigmoid, SNR=0.5, N=10^3 (b) Sigmoid, SNR=0.5, N=10^4

图 10 针对SS函数的仿真结果(大样本情形)

Fig. 10 Results for SS function under large sample cases



(a) Sigmoid, SNR=0.5, N=10^3 (b) Sigmoid, SNR=0.5, N=10^4

图 11 针对PP函数的仿真结果(大样本情形)

Fig. 11 Results for PP function under large sample cases

4.2 实际数据(Real-world cases)

本部分将利用来自不同应用领域的10个实际数据集, 进一步验证VM的有效性. 数据集来自:

1) UCI: http://archive.ics.uci.edu/ml/datasets.html.

2) Delve: http://archive.ics.uci.edu/ml/datasets.html.

10个数据集的具体信息见表1.

这里本文只考虑Sigmoid激活函数情形. 在仿真过程中, 首先按照原数据集中样本先后顺序取定训练和测试样本; 然后, 采用常用的网格搜索策略[2–5,7,20]: $L \in \{1, 2, \cdots, L_{\max}\}$($L_{\max}$视具体数据集设定), 再由5种模型选择标准进行ELM的模型选择. 在10个数据集上共进行16组实验(训练和测试样本数量见表1). 由于ELM输入参数选取的随机性, 模型选择结果有一定的波动性. 图12给出了50次模型选择得到的隐含层结点个数L和实际预测风险(测试MSE)的盒须图. 由图中结果可得以下结论:

1) 除在图12(f)和图12(i-3)所示的实验中VM性能略低于LOO外, VM都表现出更优的模型选择性能——能够选出具有最低模型复杂度(L)且实际预测风险最低(或近似最低)的ELM模型.

2) 后4个数据集上的实验结果进一步表明, VM在小样本情形下的模型选择优势更加明显.

3) 与人工数据情形类似, BIC与VM都倾向于选择预测风险低且模型复杂性低的模型, 但前者整体表现劣于后者.

4) AIC和SMS始终倾向于选择具有最多隐含层结点且实际预测风险最高的ELM模型.

表 1 回归数据集信息

Table 1 Specification of the real-world data sets

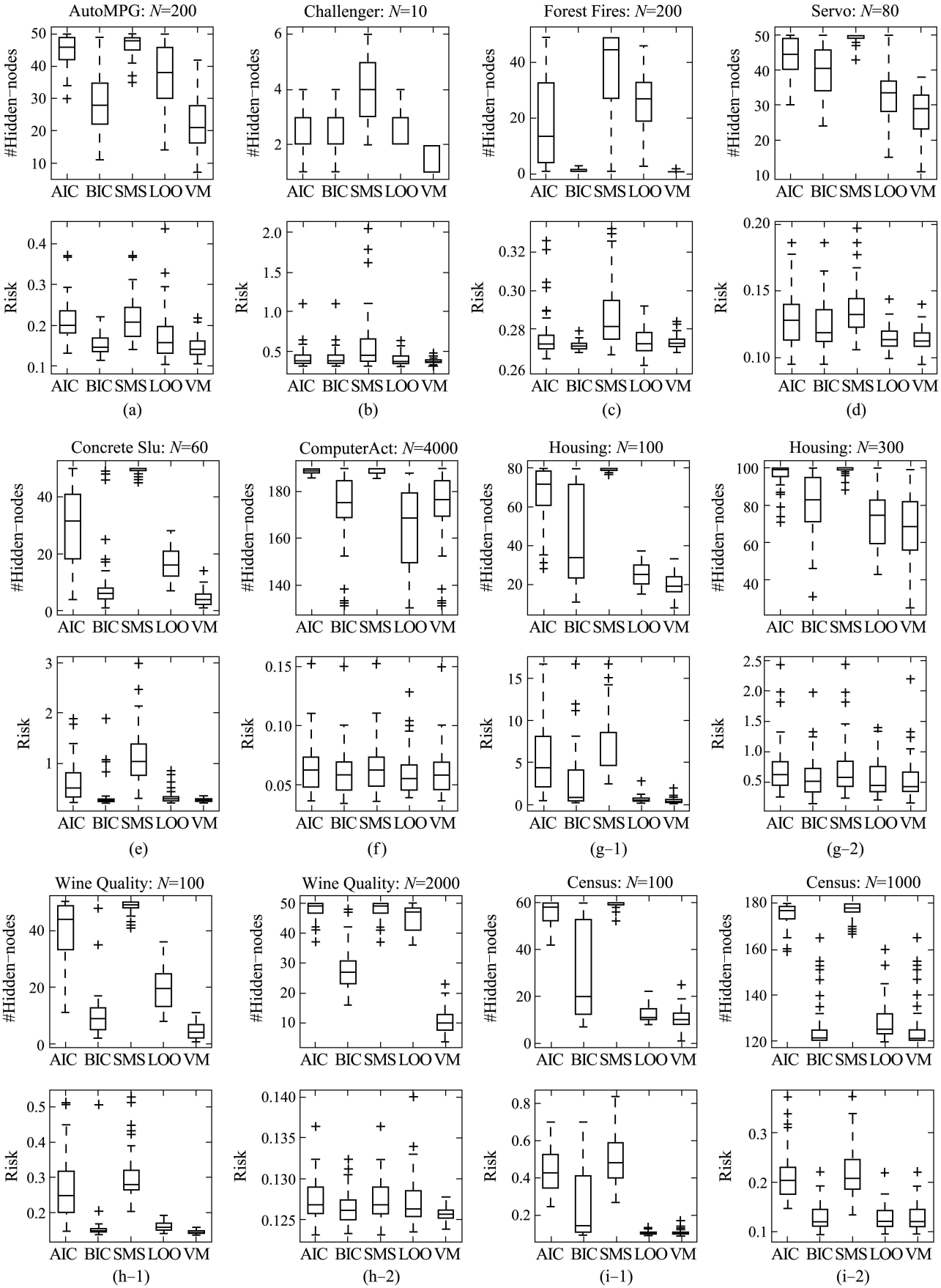
数据集	N_{train}	N_{test}	变量数	领域
Auto MPG	200	198	8	N/A
Challenger	10	13	3	Physical
Forest Fires	200	317	13	Physical
Servo	80	167	4	Computer
Concrete Slu.	60	103	10	Computer
Computer Act.	4000	4192	8	Computer
Housing	100	406	14	N/A
Housing	300	206	14	N/A
Wine Quality	100	4798	12	Business
Wine Quality	2000	2898	12	Business
Census	100	22684	8	Social
Census	1000	21784	8	Social
Census	10000	12784	8	Social
Abalone	100	4077	8	Life
Abalone	1000	3177	8	Life
Abalone	2000	2177	8	Life

VM之所以能够取得以上优势, 原因在于其能够更好地估计模型复杂性增加带来的置信风险, 并据此获得模型规模和经验风险之间的最佳折中策略. 为了更清楚地说明这一点, 图13–14分别给出了针对Abalone数据集建模过程中, 所有5种方法的实际预测风险和最优模型选择的过程信息(分$N = 100$和2000两种情况).

图13表明, 随着L的增大, VM的实际预测风险不会随着经验风险(训练MSE)一直减小, 而是在到达最优点后开始逐渐增加. 这说明VM能够对L增加所引起的置信风险进行合理的估计, 从而可以更好地控制模型复杂性. 与之相反, AIC和SMS的实际预测风险值始终与经验风险保持着很高的一致性, 都表现出持续的下降趋势. 这说明它们对模型规模引起的预测风险估计不足. 这正是AIC和SMS倾向于选择模型规模和实际预测风险都大的模型的原因. 此外, LOO的实际预测风险值表现出较大的波动性, 并倾向于选择L值较大的模型.

另一方面, 比较图13和14可知: 与小样本情形时不同, 当样本数量较大($N = 2000$)时, 实际预测风

险(测试MSE)在L的最优值(32)附近的较大范围内非常稳定; 在此情形下, VM和BIC倾向于选择该稳定范围内起始端的L值, 而AIC、SMS和LOO则相反. 这直观地解释了VM和BIC在针对中等规模样本建模时倾向于选择模型复杂性低但实际预测风险稍大的ELM模型的原因.



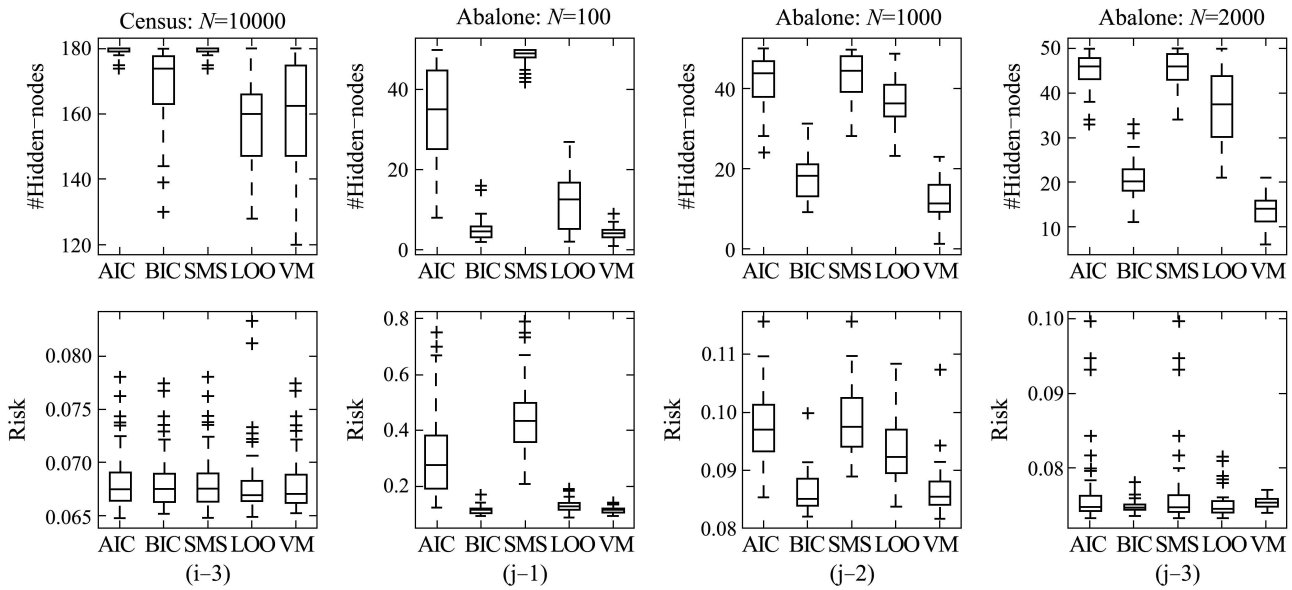


图 12 实际数据集上的比较结果

Fig. 12 Results on the real-world datasets

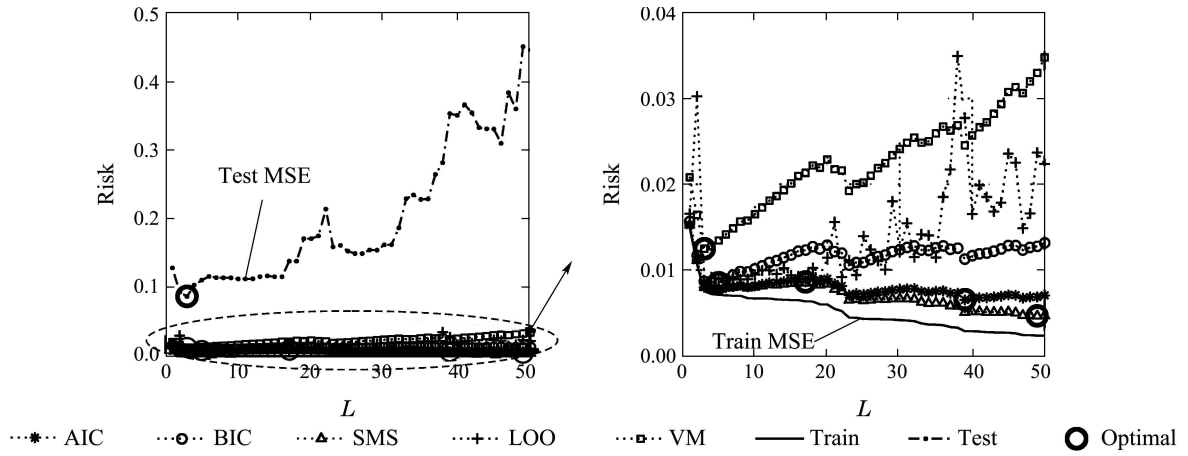


图 13 Abalone数据集上的模型选择($N = 100$)

Fig. 13 Model selection on Abalone dataset ($N = 100$)

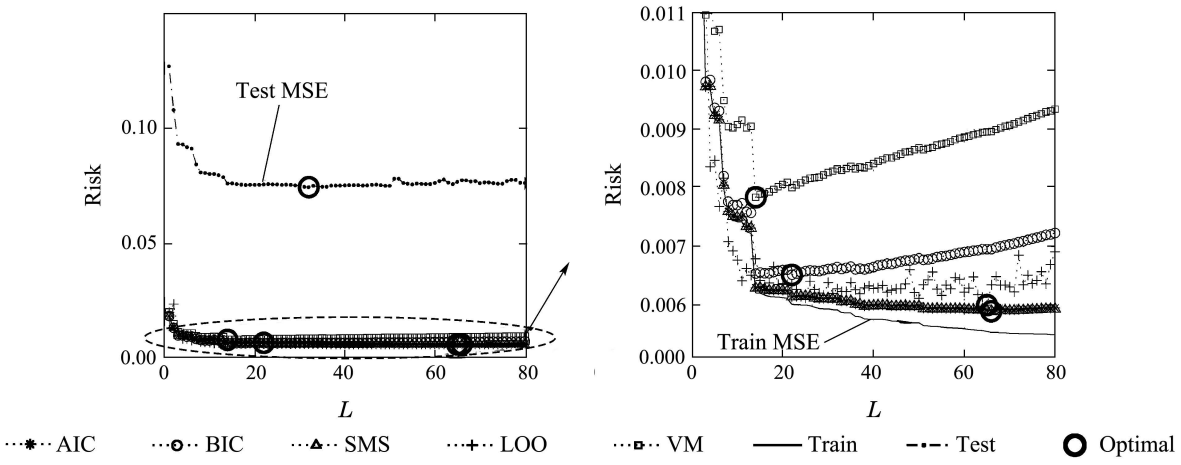


图 14 Abalone数据集上的模型选择($N = 2000$)

Fig. 14 Model selection on Abalone dataset ($N = 2000$)

5 结论(Conclusions)

对一般的学习问题, 要获得较好的泛化性能必须对模型复杂性进行合理的控制. VC泛化界给出了模型复杂性与模型泛化性能之间的关系, 具有严格

的理论基础, 且特别适合小样本情形. 本文首先讨论了基于VC泛化界的ELM模型选择方法, 然后通过在人工和实际数据集上的仿真实验, 对VM和其他4种经典模型选择方法(AIC、BIC、SMS和LOO)进行了系统的比较研究. 人工数据上的实验结果表明: VM除了在中等规模样本下性能稍差外, VM在小样本和大样本情形下都具有最佳的模型选择能力; 在10个实际数据集上的16组实验则进一步说明, 针对复杂的实际数据, VM同样具有优于其他方法的模型选择性能. 本文所得到的结果一方面为ELM模型建立过程中模型复杂性控制方法的选择提供指导, 另一方面也为VC泛化界的实际应用价值研究提供了一个例证.

参考文献(References):

[1] HAYKIN S. *Neural Networks and Learning Machines* [M]. Englewood Cliffs, NJ: Prentice-Hall, 2009.

[2] HUANG G, ZHU Q, SIEW C. Extreme learning machine: a new learning scheme of feedforward neural networks [C] //*Proceedings of International Joint Conference on Neural Networks*. Budapest, Hungary: IEEE, 2004: 25 – 29.

[3] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: theory and applications [J]. *Neurocomputing*, 2006, 70(1/2/3): 489 – 501.

[4] HUANG G B, ZHOU H, DING X, et al. Extreme learning machine for regression and multiclass classification [J]. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2011, 42(2): 513 – 529.

[5] 韩敏, 王新迎. 多元混沌时间序列的加权极端学习机预测 [J]. 控制理论与应用, 2013, 30(11): 1467 – 1472.
(HAN Min, WANG Xinying. Multivariate chaotic time series prediction based on weighted extreme learning machine [J]. *Control Theory & Applications*, 2013, 30(11): 1467 – 1472.)

[6] MICHE Y, VAN HEESWIJK M, BAS P, et al. TROP-ELM: a double-regularized ELM using LARS and Tikhonov regularization [J]. *Neurocomputing*, 2011, 74(16): 2413 – 2421.

[7] 邓万宇, 郑庆华, 陈琳. 神经网络极速学习方法研究 [J]. 计算机学报, 2010, 33(2): 279 – 287.
(DENG Wanyu, ZHENG Qinghua, CHEN Lin. Research on extreme learning of neural networks [J]. *Chinese Journal of Computers*, 2010, 33(2): 279 – 287.)

[8] RONG H J, ONG Y S, TAN A H, et al. A fast pruned-extreme learning machine for classification problem [J]. *Neurocomputing*, 2008, 72(1/2/3): 359 – 366.

[9] 尹建川, 邹早建, 徐锋. 一种基于Akaike信息准则的极限学习机 [J]. 山东大学学报(工学版), 2011, 41(6): 7 – 11.
(YIN Jianchuan, ZOU Zaojian, XU Feng. An improved extreme learning machine based on Akaike criterion [J]. *Journal of Shandong University (Engineering Science)*, 2011, 41(6): 7 – 11.)

[10] WASSERMAN L. Bayesian model selection and model averaging [J]. *Journal of Mathematical Psychology*, 2000, 44(1): 92 – 107.

[11] SHIBATA R. An optimal selection of regression variables [J]. *Biometrika*, 1981, 68(1): 45 – 54.

[12] VAPNIK V N. 统计学习理论 [M]. 许建华, 张学工, 译. 北京: 电子工业出版社, 2004.
(VAPNIK V N. *Statistical Learning Theory* [M]. XU Jianhua, ZHANG Xuegong, translated. Beijing: Publishing House of Electronics Industry, 2004.)

[13] CHERKASSKY V, MULIER F M. *Learning from Data: Concepts, Theory and Methods* [M]. New York: Wiley, 1998.

[14] CHERKASSKY V, SHAO X H, MULIER F M, et al. Model complexity control for regression using VC generalization bounds [J]. *IEEE Transactions on Neural Networks*, 1999, 10(5): 1075 – 1088.

[15] SHAO X, CHERKASSKY V, LI W. Measuring the VC-dimension using optimized experimental design [J]. *Neural Computation*, 2000, 12(8): 1969 – 1986.

[16] CHERKASSKY V, SHAO X. Signal estimation and denoising using VC-theory [J]. *Neural Networks*, 2001, 14(1): 37 – 52.

[17] CHAPELLE O, VAPNIK V, BENGIO Y. Model selection for small sample regression [J]. *Machine Learning*, 2002, 48(1): 9 – 23.

[18] MÜLLER K R, MIKA S, RÄTSCH G, et al. An introduction to kernel-based learning algorithms [J]. *IEEE Transactions on Neural Networks*, 2001, 12(2): 181 – 201.

[19] HASTIE T, TIBSHIRANI R, FRIEDMAN J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* [M]. New York: Springer-Verlag, 2001.

[20] LIU X Y, GAO C H, LI P. A comparative analysis of support vector machines and extreme learning machines [J]. *Neural Networks*, 2012, 33(9): 58 – 66.

[21] MAASS W. Neural nets with superlinear VC-dimension [J]. *Neural Computation*, 1994, 6(5): 877 – 884.

[22] CARTER M A, OXLEY M E. Evaluating the Vapnik-Chervonenkis dimension of artificial neural networks using the Poincare polynomial [J]. *Neural Networks*, 1999, 12(3): 403 – 408.

[23] BARTLETT P L, MAIOROV V, MEIR R. Almost linear VC-dimension bounds for piecewise polynomial networks [J]. *Neural Computation*, 1998, 10(8): 2159 – 2173.

[24] KARPINSKI M, MACINTYRE A. Polynomial bounds for VC dimension of sigmoidal and general pfaffian neural networks [J]. *Journal of Computer and System Science*, 1997, 54(1): 169 – 176.

[25] HUANG G B, CHEN L, SIEW C K. Universal approximation using incremental constructive feedforward networks with random hidden nodes [J]. *IEEE Transactions on Neural Networks*, 2006, 17(4): 879 – 892.

[26] HUANG G B, CHEN L. Convex incremental extreme learning machine [J]. *Neurocomputing*, 2007, 70(16): 3056 – 3062.

[27] HUANG G B, CHEN L. Enhanced random search based incremental extreme learning machine [J]. *Neurocomputing*, 2008, 71(16): 3460 – 3468.

[28] HARDLE W, HALL P, MARRON J S. How far are automatically chosen regression smoothing parameters from their optimum [J]. *Journal of the American Statistical Association*, 1988, 83(401): 86 – 95.

[29] AKAIKE H. Fitting autoregressive models for prediction [J]. *Annals of the Institute of Statistical Mathematics*, 1969, 21(1): 243 – 247.

[30] CHERKASSKY V, MA Y. Comparison of model selection for regression [J]. *Neural Computation*, 2003, 15(7): 1691 – 1714.

[31] 刘学艺, 李平, 郜传厚. 极限学习机的快速留一交叉验证算法 [J]. 上海交通大学学报, 2011, 45(8): 1140 – 1145.
(LIU Xueyi, LI Ping, GAO Chuanhou. Fast leave-one-out cross-validation algorithm for extreme learning machine [J]. *Journal of Shanghai Jiaotong University*, 2011, 45(8): 1140 – 1145.)

作者简介:

刘学艺　(1978–), 男, 博士研究生, 主要研究方向为机器学习、人工智能以及复杂工业过程建模与优化等, E-mail:zjuliuxy@163.com;

宋春跃　(1972–), 男, 博士, 教授, 主要研究方向为混杂系统优化控制、生产调度以及随机系统优化控制等, E-mail: cysong@iipc.zju.edu.cn;

李　平　(1954–), 男, 博士, 教授, 主要研究方向为复杂工业过程建模与控制、混杂系统理论以及控制与导航等, E-mail: pli@iipc.zju.edu.cn.