

训练误差概率分布对泛化性能的影响

赵金伟, 王宇飞, 柳宇, 黑新宏

(西安理工大学计算机科学与工程学院, 西安 710048)

摘要: 当损失函数确定后, 在经验风险中平等对待给定的样本数据集是不正确的, 因此在分析无界损失函数集合的风险界时, 必须考虑训练样本预测误差概率分布的尾部厚重程度。本文将训练样本的预测误差按倒序排序, 然后重新讨论经验风险与结构风险, 分析期望风险的上界, 提出新的泛化误差界——尾部综合指数下界。基于商空间理论, 提出将尾部综合指数下界与留一法 (leave-one-out, LOO) 误差界综合考虑, 在假设空间构造多层拓扑结构来寻找最优的假设紧子空间。并通过实验证明了对于一些学习算法, 本方法可以提高其泛化性能的鲁棒性和有效性。

关键词: 人工智能; 无界非负损失函数集; 统计学习理论; 拓扑结构; 重尾指数

中图分类号: TP18 **文献标识码:** A **文章编号:** 1674-2850(2018)16-1586-13

Influence of training error probability distribution on generalization performance

ZHAO Jinwei, WANG Yufei, LIU Yu, HEI Xinhong

(School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: When the loss function is determined, it is not correct to treat the given sample data set equally in empirical risk. Therefore, when analyzing the risk bound of the set of unbounded loss functions, the thickness of tail in the training sample prediction error probability distribution must be considered. In this paper, the prediction errors of the training sample are sorted in ascending order, then new empirical risk and structural risk are discussed, and the upper bound of the expected risk is analyzed. A new generalization error boundary, lower bound of tail synthetic index, is proposed. Based on quotient space theory, the new bound of tail synthetic index and leave-one-out (LOO) error bound are introduced for finding an optimal compact hypothesis subspace by constructing a multi-layer topology in hypothesis space. The robustness and effectiveness of the proposed method in improving the generalization performance of some learning algorithms are validated by experiments.

Key words: artificial intelligence; unbounded nonnegative loss function set; statistical learning theory; topological structure; tail synthetic index

0 引言

利用有限样本获得强泛化性能学习模型是机器学习研究的终极目标。如何对候选模型进行评估也是模型选择的重要依据^[1]。20世纪60年代, CHERKASSKY^[2]提出了统计学习理论, 给出了训练样本预测误差分布即训练误差分布的峰度、VC维及有界损失函数集合上界等泛化误差界, 并以此建立了假设空间的拓扑结构来缩小搜索空间, 然而这些界都是非构造的。CHERKASSKY^[2]又提出几何间隔及不依赖空

基金项目: 国家自然科学基金 (61672027, 61773314, 61773313); 高等学校博士学科点专项科研基金 (20136118120011); 陕西省自然科学基金基础研究计划 (2017JM6080); 国家重点研发计划 (2017YFB1201500)

作者简介: 赵金伟 (1974—), 男, 副教授, 主要研究方向: 机器学习

通信联系人: 黑新宏, 教授, 主要研究方向: 轨道交通、机器学习、数据分析. E-mail: heixinhong@xaut.edu.cn

间维数上界的 fat 维替代 VC 维。之后，人们又提出了 Radius Margin 界^[3-4]、Span 界^[5]、GCV 界^[6]、LSSVM 误差界^[7]、基于 Rademacher 复杂度的泛化误差界^[8]、基于最大差异估计的泛化误差界^[9]、局部复杂度的泛化误差界^[10]等作为学习机器泛化性能的评价准则。DEVROYE 等^[11]研究了留一稳定性，提出了最近邻模型泛化性能的度量指标。BOUSQUET 等^[12]研究了假设稳定性、逐点假设稳定性和一致稳定性，分析了正则学习框架相关算法的稳定性与泛化性能之间的关系。POGGIO 等^[13]为深入研究算法稳定性与泛化性能之间的关系，总结并发展了稳定性学习理论。DING 等^[14]提出了再生核希尔伯特空间 (RKHS) 中的球覆盖数评估模型的泛化性能。又有很多学者给出了核正则化学习框架下期望误差的上界^[15-18]。目前已有工作证明，利用核方法解决大数据问题时的精度与深度神经网络的预测精度相近^[19]，而且模型更简单，更易于解释。冯昌等^[20]基于高斯核函数的随机傅里叶特征近似，提出了一种新的、高效的核支持向量机模型选择方法，将无限维隐式特征空间嵌入到一个相对低维的显式随机特征空间，并推导了在两个不同的特征空间中分别训练所得到模型的误差上界。针对大数据分析问题，冯昌等^[21]进一步给出了基于大规模核方法的循环随机假设空间算法的一致泛化误差界及其相对于最优泛化误差的收敛性。

然而在依据上述泛化误差界评价模型的泛化性能时，并未考虑训练样本预测误差概率分布“尾部”特征对学习机器泛化性能的影响。实际上，VAPNIK 在文献^[4]中已讨论并指出，为得到更精细的风险界，必须掌握某些关于训练样本预测误差分布“尾部”的先验信息。因此，本文讨论训练样本预测误差概率分布“尾部”特征与模型泛化性能之间的关系，并重新推导期望风险上界。基于极值理论，提出尾部综合指数下界，并在模型选择中同时利用尾部综合指数下界和 LOO 误差界对学习机器泛化性能进行全面评价。将该方法应用到多种最小二乘支持向量机 (least squares support vector machine, LSSVM) 算法中，并与传统算法进行性能比较。

1 算法理论基础

1.1 损失函数空间的风险界分析

统计学习理论指出，当容许损失函数集 $Q(z, \alpha)$, $\alpha \in A$, A 为有界损失函数集 ($Q(z, \alpha) < B$, B 为有界损失函数集 A 的上界) 时，

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{B\varepsilon(\ell)}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha)}{B\varepsilon(\ell)}} \right) \quad (1)$$

依至少 $1-\eta$ 的概率成立^[2]，其中， $0 \leq \eta \leq 1$, $\varepsilon(\ell) = 4 \frac{G^{A,B}(2\ell) - \ln(\eta/4)}{\ell}$, $G^{A,B}(2\ell)$ 为生长函数， ℓ 为样

本个数；期望风险 $R(\alpha) = \int Q(z, \alpha) dF(z)$, $F(z)$ 为样本的分布函数；经验风险 $R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)$, $i = 1, 2, \dots, \ell$.

而对于容许损失函数集为无界损失函数集的所有函数来说，

$$R(\alpha) \leq \left(\frac{R_{\text{emp}}(\alpha)}{1 - \tau a(p) \sqrt{\varepsilon_2(\ell)}} \right)_{\infty}, \quad a(p) = \sqrt[p]{\frac{1}{2} \left(\frac{p-1}{p-2} \right)^{p-1}} \quad (2)$$

依至少 $1-\eta$ 的概率成立，其中， $\tau = \sup_{\alpha, z} \frac{\sqrt[p]{EQ^p(z, \alpha)}}{EQ(z, \alpha)}$ 为整体样本空间预测误差的峰度， $\sqrt[p]{EQ^p(z, \alpha)}$ 为 $Q(z, \alpha)$ 的 p 阶矩， p 为整数。

不等式(1)和(2)的右端代表结构风险。如果提供一个容许结构,然后在整个给定的结构元素即容许函数集 \mathcal{F} 上就可找到最小化保证风险[即式(1)和式(2)]的函数 f ,其中,容许结构是依据VC维 h 、有界损失函数集的上界 B 或无界损失函数集的峰度 τ 等几个影响泛化性能的重要因素的大小次序,将函数空间 E 分割成一个个由函数子集 $S_1 \subset S_2 \subset \dots \subset S_n$ 构成的函数空间拓扑结构。

文献[3]指出,经验风险是训练样本预测误差的平均值。然而,当函数确定后,对于给定的样本数据集来说,训练样本的预测误差在整体样本空间中并不全是大概率事件[22],这时在经验风险中平等对待这些样本是不正确的,因此在分析风险界时,应考虑训练误差概率分布的尾部厚度。由统计学理论可知,峰度 τ 是描述概率分布尾部厚重程度的理想统计量,然而它只适用于高价值数据,对于低价值甚至小样本数据集来说峰度 τ 并不合适。为此,首先将这些训练样本的预测误差按其大小进行排序,然后重新讨论经验风险与结构风险。

以无界非负损失函数集中的任意损失函数 $Q(z, \alpha)$ 为例,将样本依据其损失函数 $Q(z, \alpha)$ (即预测误差的大小)排序,得到一个正序序列 $z_1, z_2, \dots, z_i, \dots, z_\ell$,再以 $Q(z_k, \alpha)$ 为阈值,将上述序列分为两部分,一个是非极端值部分 $(z_1, \dots, z_i, \dots, z_{k-1})$,另一个是极端值部分 $(z_k, \dots, z_i, \dots, z_\ell)$,前者描述了误差分布中的大概率情况,后者描述了小概率情况。下面从定理1来讨论无界非负损失函数集中一般情况下的结构风险问题。

定理1 设无界非负损失函数集为 $Q(z, \alpha)$, $\alpha \in \Lambda$, $F(z)$ 为样本 z 的分布函数, $H_{\text{ann}}^S(2\ell)$ 为函数集合 \mathcal{F} 在 ℓ 个样本中的退火熵,则不等式

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\int Q(z, \alpha) dF(z)} > \varepsilon \right\} < 4 \exp \left\{ \left[\frac{H_{\text{ann}}^S(2(k-1))}{k} - \frac{\left[\varepsilon - \frac{1}{2} + \min_{\alpha \in \Lambda} \left(\frac{\chi}{\ell} \right) \right]^2}{4(2k-1)} \right] k \right\} \quad (3)$$

成立,其中, $\chi = Q(z_{k-1}, \alpha) \sum_{i=k}^{\ell} [\ln Q(z_i, \alpha) - \ln Q(z_{k-1}, \alpha)]$ 。

令不等式(3)的右端等于 η ,可得

$$\varepsilon = \frac{1}{2} - \frac{\min_{\alpha \in \Lambda}(\chi)}{\ell} + \sqrt{4(2k-1) \left[\frac{h \ln \frac{2(k-1)}{h} + h}{k} - \frac{1}{k} \ln \frac{\eta}{4} \right]} \quad (4)$$

由式(3)可进一步推出

$$R(\alpha_\ell) \leq R_{\text{emp}}(\alpha_\ell) \left\{ \frac{1}{2} + \frac{\min_{\alpha \in \Lambda}(\chi)}{\ell} - \sqrt{4(2k-1) \left[\frac{h \ln \frac{2(k-1)}{h} + h}{k} - \frac{1}{k} \ln \frac{\eta}{4} \right]} \right\}^{-1} \quad (5)$$

依 $1-\eta$ 的概率成立,其中, $\chi = Q(z_{k-1}, \alpha) \sum_{i=k}^{\ell} [\ln Q(z_i, \alpha) - \ln Q(z_{k-1}, \alpha)]$ 称为尾部综合指数。

由式(5)可以看出,当训练样本集合和函数集合 \mathcal{F} 确定后,风险界的大小与尾部综合指数 χ 的下界有关。当损失函数集合的尾部综合指数 χ 的下界上升时,可以引起不等式右边的分母增大,使得置信

区间减小。但是当尾部综合指数 χ 下界太大时, 函数集合又会太小, 使得经验风险变大。这说明, 损失函数集合的尾部综合指数的下界影响着学习机器的泛化性能。

1.2 尾部综合指数

本节将重点讨论如何利用训练样本的预测误差计算尾部综合指数 χ , 以全面了解其对模型泛化性能的影响。

由于损失函数的尾部综合指数与训练样本预测误差的极端值部分相关, 它描述了误差分布的尾部特征。假设误差分布服从正态分布, 而正态分布和轻尾分布都可以看成是重尾分布的一种特殊情况。依据极值分布类型定理^[23]可知, 对于一个具有重尾分布特点独立同分布随机变量序列, 其尾部厚薄程度和趋近于零的速度, 称为尾部指数 a , 亦称重尾指数^[24~25]。重尾指数越小, 则尾部越厚, 趋近于零的速度越慢; 反之亦然。

在极值理论中常假设样本数趋于无穷大以进行极值分布的参数估计, 然而实际中这种假设是不能满足的。可依据 Hill 估计法^[26], 即式 (6) 求出重尾指数 a 的值。

$$a^{-1} = \frac{1}{m} \sum_{i=1}^m \log X_n^{(i)} - \log X_n^{(m+1)}. \quad (6)$$

由上述讨论可知, 若要清楚地描述机器学习过程中训练样本预测误差的尾部分布情况, 只需正确估计出其重尾指数 a 与所有极端值点的个数 m 就可以了。由 1.1 节讨论和式 (6) 可知, 尾部综合指数 χ 可由重尾指数 a 与极端值点的个数 m 得到, 其计算公式为

$$\chi = m/a X_n^{(m+1)}, \quad (7)$$

其中, 极端值点个数 $m = \ell - k$, k 为非极端值训练样本点个数。对式 (7) 进行如下讨论: 当要求的尾部综合指数 χ 增大时, 通常会引起非极端值点个数 k 减小, 同时也会引起重尾指数 a 减小。由 1.1 节中的风险界 (5) 可知, 当非极端值点个数 k 减小时, 随着 χ 下界的增大, 式 (5) 不等号右边的分母将增大, 从而使得置信区间减小。由重尾指数的定义^[27]可知, 当重尾指数 a 减小时, 会引起训练误差分布的尾部变厚, 训练样本的预测误差更集中于误差分布的尾部区域, 经验风险的计算更不准确, 从而降低算法的泛化性能。而当尾部综合指数 χ 减小时, 正好相反。因此, 在学习过程中, 尾部综合指数能够综合考虑置信区间与经验风险的变化情况。

1.3 构造多层拓扑结构中涉及的基本问题

通过实验发现, 对于无界非负损失函数集合来说, 学习机器的泛化性能还与尾部综合指数下界有关。如何利用尾部综合指数下界选择更好的模型是本节需要解决的问题。

本文基于尾部综合指数下界和传统的泛化误差界共同确定一个假设紧子集。首先利用上述不同的广义泛化误差界^[28~29]对假设空间构建不同的商空间, 然后在这些商空间的交集中寻找能保证结构风险最小的假设紧子空间, 从中搜索最优模型。其中涉及以下几个问题: 第一, 如何利用多种广义泛化误差界对假设空间进行多层拓扑结构构造; 第二, 如何提高多层拓扑结构的构建效率; 第三, 如何确定每层商空间粒度。

1.3.1 假设空间中多层拓扑结构的构造

基于以下几个定义, 在假设空间中构造多层拓扑结构。

定义 1 (模糊等价关系) 给定一个训练数据集, 在一个函数集合 \mathcal{F} 中, 存在评价函数 ϕ , 使得任意两个函数 f_1 和 f_2 的评价结果分别为 τ_1 和 τ_2 , 则称 $R(f_1, f_2) = 1 - |\tau_1 - \tau_2|$ 为这两个函数 f_1 和 f_2 之间的模糊等价关系。

定义 2 (普通等价关系和截断值) 设 R 为函数集合 \mathcal{F} 上的模糊等价关系, 令 $R_\lambda = \{(f_1, f_2) | R(f_1, f_2) \geq \lambda\}$, $0 \leq \lambda \leq 1$, 则 R_λ 为函数集合 \mathcal{F} 上的普通等价关系。其中, λ 为截断值。

定义 3 (等价类) 若一个函数子集 \mathcal{M} 中的任意两个函数 f_1 和 f_2 之间都存在普通等价关系 R_λ , 就称这个函数子集为一个等价类, 也称为一个粒子。

定义 4 (商空间) 若函数集合 \mathcal{F} 有一个依据普通等价关系 R_λ 所得到的等价类子集 $\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_i, \dots\}$, 其中任意两个等价类 \mathcal{M}_i 和 \mathcal{M}_j 不相交, 且 $\mathcal{F} = \bigcup_i \mathcal{M}_i$, 就称这个等价类子集 $\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_i, \dots\}$ 构成了函数集合 \mathcal{F} 的一个商空间 $[\mathcal{F}]$ 。

定义 5 (粒度) 设 $\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_i, \dots\}$ 为商空间 $[\mathcal{F}]$ 中的所有等价类集合, 函数集合 $\mathcal{F} = \bigcup_i \mathcal{M}_i$, 则称粒子 \mathcal{M}_i 的粒度为 $d(\mathcal{M}_i) = |\mathcal{M}_i| = \int_{x \in \mathcal{M}_i} dx$ 。

由粒度和截断值的定义可知, 当截断值减小时粒度就越大, 称粒子越粗; 反之, 粒度越小, 称粒子越细。

定义 6 (多层拓扑结构) 若有多个普通等价关系 $R_\lambda, R'_\lambda, R''_\lambda, \dots$, 首先利用 R_λ 对函数空间 \mathcal{F} 进行划分, 得到商空间 $[\mathcal{F}]$, 再利用 R'_λ 对该商空间中的每一个粒子进行划分, 就会得到下层的商空间 $[\mathcal{F}]'$, 依次类推, 就会得到多个层次的商空间 $\{[\mathcal{F}], [\mathcal{F}]', [\mathcal{F}]'', \dots\}$, 将这样的结构称为多层拓扑结构。

然而, 在利用该方法构造多层拓扑结构时, 还存在以下问题需要解决: 1) 假设空间是无限集合, 实际上不可能穷尽所有函数来计算每一种泛化误差界的评价函数值, 那么如何高效地构造多层拓扑结构呢? 2) 如何确定每一种泛化误差界的截断值, 即如何确定每一层拓扑结构的粒度?

1.3.2 提高构建多层拓扑结构效率的方法

因为假设空间是无限的, 所以如何高效地构建多层拓扑结构是一个难点。本课题组认为, 解决问题的关键应是如何对学习过程中所搜索到的函数进行多层拓扑结构的构建。因此, 本节给出一种自顶向下构建多层拓扑结构的方法。

若给定训练样本集合, 针对当前搜索得到的函数, 首先计算出第一个评价函数值, 然后, 利用事先给定的第一个截断值, 获得该函数在第一层商空间中所属的等价类。利用同样的办法计算出该函数在以下各层商空间中所属的等价类。随着搜索的进行, 最终将会得到一个由有限的已被搜索函数的集合构成的多层拓扑结构^[30]。

缩减每层商空间上的等价类个数可以提升结构的构建效率。本文利用启发式方法, 保留每层商空间上的一些等价类, 而舍弃其他等价类。以 LOO 泛化误差界作为评价函数为例, 在构建拓扑结构时, 只需保留平均 LOO 泛化误差界最小的那些等价类即可。而对于其他评价函数, 可指定保留多个等价类。

由 1.1 节和 1.2 节的讨论可知, 当尾部综合指数 χ 的下界上升时, 风险界的置信区间会减小, 但是当尾部综合指数 χ 的下界上升到一定值时, 又会引起函数集合缩小, 使得经验风险变大。因此, 尾部综合

指数的下界不应太大也不应太小。这一点将在 3.2 节通过实验进行证明。总之, 基于尾部综合指数下界决定等价类的舍弃, 需要确定尾部综合指数的上下界。

1.3.3 每层商空间粒度的确定

尾部综合指数上下界与等价类的粒度密切相关, 而且较小的粒度有助于进一步提高拓扑结构构建效率。由 1.3.1 节的粒度定义可知, 每层商空间的粒度^[31]与普通等价关系的截断值有关。由于较大的尾部综合指数的上下界之差能够保证较小的经验风险, 所以尾部综合指数的截断值应小于 1 且不能等于 1; 另外, 还应注意截断值不能等于 0, 否则将失去意义。因此, 本文首先依据上述条件确定截断值, 然后利用启发式方法来确定尾部综合指数的上下界。对于 LOO 泛化误差界来说, 由于评价函数就是其本身, 对于每个粒子来说不存在上下界的问题, 所以粒子中只包含具有相同 LOO 泛化误差界的函数, 故可以令它的截断值等于 1。

2 基于尾部综合指数下界的模型选择方法

本节重点介绍在 LSSVM 算法中, 如何应用假设空间多层拓扑结构方法进行模型选择。利用尾部综合指数下界和 LOO 泛化误差界来构造两层拓扑结构。该方法首先由预设定的截断值和尾部综合指数上下界计算第一层商空间中等价类的个数和每个等价类尾部综合指数的下界; 然后利用遗传算法或粒子群算法搜索得到的当前核参数和超参数进行学习预测, 计算当前回归函数的尾部综合指数, 并判断当前参数所决定的回归函数所在的等价类^[32~33]。若所在的等价类属于所指定的等价类范围, 则利用预测误差处在非极端点的样本再次进行学习, 计算 LOO 泛化误差界, 并设置优化目标值为当前 LOO 泛化误差界, 否则设置优化目标值为最大。每次迭代, 对于第一层商空间中的每一个等价类, 只保留 LOO 泛化误差界最小的那一组参数, 迭代结束后返回。最终对各个等价类中的最优参数依据择优策略进行选择。实际上不同的择优策略会得到不同的寻参方法, 本文以 LOO 泛化误差界最小的那组参数为最优参数。而利用优化目标值寻参的策略, 可以采用粒子群算法或遗传算法^[34~35]。学习阶段, 首先利用这组最优参数, 进行学习和预测, 计算出训练样本预测误差, 利用式 (6) 和式 (7), 计算重尾指数 a 、最优 k 值和尾部综合指数 χ 。统计学习理论指出, 在利用经验风险最小化方法搜索最好的保证解的过程中, 可以试图排除训练数据的某一子集, 以得到最好的界^[2]。然后, 将预测误差中处于尾部的样本点剔除后, 利用其余的样本点再次进行学习, 最终得到回归模型, 学习结束。

模型选择阶段, 计算优化目标值中, 比通常的方法多进行一次学习与预测, 时间复杂度为 $O(\ell^2)$ 。计算尾部综合指数的时间复杂度为 $O(\ell)$ 。当尾部综合指数处于指定的范围内时, 进行下一步的计算优化目标值的过程, 时间复杂度为 $O((\ell-k)^2)$; 若不在指定范围内时, 不计算优化目标值, 因此整个程序的时间复杂度为 $O(\ell^2)$ 。

学习阶段, 预先学习的时间复杂度为 $O(\ell^2)$, 计算尾部综合指数的时间复杂度为 $O(\ell)$, 学习的时间复杂度为 $O((\ell-k)^2)$, 故整个学习过程的时间复杂度为 $O(\ell^2)$ 。

由上述分析可以发现, 本算法对寻参与学习过程的时间复杂度影响不大。

3 实验验证与讨论

本节以 UCI^[36]标准数据集中的 Cadata、Cpussmall、Housing、Mpg、Triazines、Mg 和 Space_ga 数据集

为例,测试了将多层拓扑结构构造方法应用于 LSSVM^[37]和加权最小二乘支持向量机(RobustLSSVM)^[38-40]中时,对算法泛化性能的影响。其中测试环境为 Matlab R2009a,测试时使用 LSSVM 工具箱 LSSVmlabv1_8_R2006a_R2009a 和高斯核函数。模型泛化性能评价指标为预测均方误差(mean square error, MSE)和相关系数(spearman correlation coefficient, SCC)。

3.1 标准数据集

实验采用 UCI^[36]的 Cadata、Cpusmall、Housing、Mpg、Triazines、Mg 和 Space_ga 标准回归数据集。在每个数据集中随机抽取了 100 个样本构成数据集。从 Cadata、Cpusmall、Space_ga、Mg 中随机抽取 10 次,生成 10 个样本集。从 Housing 中随机抽取 5 次,从 Mpg 中随机抽取 4 次,从 Triazines 中随机抽取 2 次。然后采用 holdout 方法将每个数据集样本的 70%作为训练集,剩下的 30%作为验证集。另外,为消除数据各分量不同量纲之间的影响和提高算法的收敛速度,在学习前,需要对这些数据进行归一化处理。实验中所用的标准数据集的特征与划分情况如表 1 所示。

表 1 各种标准数据集

Tab. 1 Bench mark data sets

数据集	属性个数	样本个数
Cadata-1,⋯, Cadata-10	8	100
Cpusmall-1,⋯,Cpusmall-11	12	100
Housing-1,⋯, Housing-5	13	100
Mpg-1,⋯, Mpg-4	7	100
Triazines-1,⋯, Triazines-2	60	100
Mg-1,⋯, Mg-10	6	100
Space_ga-1,⋯,Space_ga-10	6	100

3.2 算法必要性验证

本节通过实验验证将尾部综合指数引入模型选择的必要性,以及尾部综合指数下界与截断值的设置依据。本实验以 LOO 泛化误差界为优化目标,分别计算出待选模型的尾部综合指数、LOO 泛化误差界和测试样本预测结果的均方误差等三个指标,然后讨论它们之间的关系。其惩罚因子与核参数分别在 (0.01,100) 之间利用网格搜索法取值。以 LOO 泛化误差界为纵坐标,以尾部综合指数为横坐标,以 MSE 为色度,在一个二维坐标系中,画出相应的彩色点。由于篇幅的原因,下面只给出典型结果,即数据集 Cadata 的实验结果,如图 1 所示。

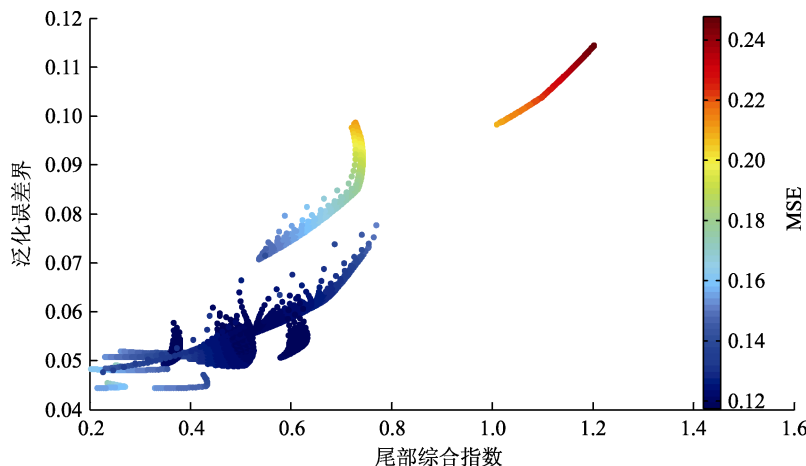


图 1 测试样本预测结果的 MSE、LOO 泛化误差界和尾部综合指数之间的变化关系

Fig. 1 Relationship among MSE, LOO generalized error bond and the tail composite index in the prediction result of test samples

图 1 中每个彩色点为一个三元组,这三个元分别为 LOO 泛化误差界、尾部综合指数和 MSE. 在图 1 中,若保持尾部综合指数不变,可以发现,当 LOO 泛化误差界最小时, MSE 不一定能保证最小,相反,还会变大。若保持 LOO 泛化误差界最小且不变时,发现较深的点(即算法泛化性能较好的点)往往集

中在尾部综合指数指示的某个区间内。同时, 这些区间内颜色最深的点, 其 LOO 泛化误差界往往也保持最小。这些现象表明, 在进行模型选择时, 只考虑 LOO 泛化误差界最小, 往往是不准确的。但如果引入尾部综合指数将函数空间划分为一个个粒子, 然后在最合适的粒子中寻找最优回归模型, 往往会取得比较好的效果。这就证明在模型选择与学习阶段引入尾部综合指数进行函数集合的粒划分是正确的。由于尾部综合指数可以很方便地通过训练样本的预测误差得到, 是可构造的, 故这种方法是可行的。由图 1 还可以看出, 当用尾部综合指数对函数集合进行粒划分时, 这些粒子的粒度不能太小, 故尾部综合指数的截断值应该取 0~1 之间的值。

3.3 算法泛化性能测试

本节将尾部综合指数引入模型选择后的 LSSVM 算法和 RobustLSSVM 算法, 即 QT-LOOLSSVM 和 QT-RobustLSSVM, 与未应用该方法的这两种算法进行比较。其中标准数据集的准备过程如 3.1 节所述。利用 holdout 方法将每个数据集样本的 70% 作为训练集进行学习训练, 其余 30% 作为验证集进行模型的预测和验证。每组实验重复 10 次, 计算出 MSE 和 SCC 的平均值, 其结果如表 2 和表 3 所示。本实验采用 Nelder-Mead simplex optimization algorithm 算法^[41]和 coupled simulated annealing 算法^[42]作为寻参策略。

表 2 本方法在 LSSVM 算法上的泛化性能比较

Tab. 2 Comparison of the generalization performance of the proposed method on LSSVM algorithm

数据集	SV	χ_{\min}	d	算法			
				QT-LOOLSSVM		LOOLSSVM	
				MSE	SCC	MSE	SCC
Cadata-1	95	0.3	0.2	0.267 032	0.578 401	0.271 758	0.487 281
Cadata-2	92	0.3	0.2	0.069 969	0.061 171	0.102 426	0.002 833
Cadata-3	93	0.3	0.2	0.052 688	0.090 332	0.102 963	0.009 691
Cadata-4	95	0.3	0.2	0.134 260	0.465 270	0.229 141	0.018 974
Cadata-5	98	0.3	0.2	0.041 722	0.355 579	0.045 043	0.293 341
Cadata-6	96	0.3	0.2	0.064 347	0.187 656	0.145 062	0.010 086
Cadata-7	94	0.3	0.2	0.098 245	0.594 902	0.142 858	0.144 375
Cadata-8	93	0.3	0.2	0.059 905	0.794 685	0.211 129	0.161 481
Cadata-9	93	0.3	0.2	0.061 598	0.333 843	0.130 954	0.089 237
Cadata-10	96	0.3	0.2	0.153 620	0.069 702	0.219 230	0.046 391
Cpusmall-1	97	0.3	0.2	0.005 364	0.775 935	0.009 848	0.785 910
Cpusmall-2	94	0.3	0.2	0.008 211	0.657 455	0.016 308	0.609 412
Cpusmall-3	98	0.3	0.2	0.005 353	0.794 969	0.011 436	0.611 684
Cpusmall-4	98	0.3	0.2	0.005 682	0.844 548	0.008 012	0.780 248
Cpusmall-5	94	0.3	0.2	0.006 096	0.757 288	0.031 343	0.302 747
Cpusmall-6	97	0.3	0.2	0.005 739	0.790 281	0.008 313	0.704 594
Cpusmall-7	94	0.3	0.2	0.210 073	0.387 741	0.253 459	0.256 544
Cpusmall-8	94	0.3	0.2	0.126 539	0.305 126	0.209 272	0.075 241
Cpusmall-9	97	0.3	0.2	0.019 740	0.926 102	0.053 137	0.603 275
Cpusmall-10	97	0.3	0.2	0.009 338	0.757 328	0.027 466	0.593 215
Cpusmall-11	96	0.3	0.2	0.016 830	0.822 684	0.083 758	0.657 175
Housing-1	98	0.3	0.2	0.017 713	0.849 397	0.037 897	0.692 450
Housing-2	91	0.3	0.2	0.111 872	0.676 057	0.362 279	0.168 575
Housing-3	92	0.3	0.2	0.040 538	0.849 069	0.047 425	0.792 852

续表

数据集	SV	χ_{\min}	d	算法			
				QT-LOOLSSVM		LOOLSSVM	
				MSE	SCC	MSE	SCC
Housing-4	92	0.3	0.2	0.301 659	0.352 855	0.320 735	0.810 797
Housing-5	95	0.3	0.2	0.146 269	0.152 038	0.211 147	0.002 017
Mpg-1	93	0.3	0.2	0.025 039	0.846 718	0.027 865	0.826 356
Mpg-2	93	0.3	0.2	0.064 177	0.707 974	0.048 971	0.802 323
Mpg-3	96	0.3	0.2	0.028 753	0.850 349	0.028 234	0.880 162
Mpg-4	96	0.3	0.2	0.092 035	0.508 420	0.100 605	0.546 432
Triazines-1	96	0.3	0.2	0.195 272	0.184 649	0.208 164	0.127 012
Triazines-2	97	0.3	0.2	0.214 312	0.112 723	0.208 164	0.127 012
Mg-1	90	0.5	0.2	0.287 609	0.176 358	0.364 513	0.106 986
Mg-2	83	0.3	0.2	0.045 751	0.856 689	0.067 980	0.794 784
Mg-3	95	0.3	0.2	0.148 981	0.504 588	0.150 151	0.501 355
Mg-4	89	0.7	0.3	0.143 990	0.511 398	0.126 141	0.531 863
Mg-5	94	0.5	0.2	0.074 324	0.790 749	0.081 234	0.713 574
Mg-6	91	0.5	0.2	0.080 802	0.693 866	0.102 096	0.620 390
Mg-7	91	0.5	0.2	0.049 563	0.836 823	0.089 452	0.698 455
Mg-8	87	0.5	0.2	0.063 268	0.783 620	0.107 862	0.614 268
Mg-9	89	0.5	0.2	0.098 727	0.654 503	0.125 701	0.599 558
Mg-10	93	0.5	0.2	0.069 967	0.739 756	0.057 292	0.790 561
Space_ga-1	92	0.3	0.2	0.180 221	0.156 604	1.038 397	0.123 278
Space_ga-2	92	0.3	0.2	0.376 966	0.177 802	40.621 65	0.184 540
Space_ga-3	91	0.3	0.2	0.016 312	0.877 293	0.025 124	0.818 067
Space_ga-4	93	0.3	0.2	0.015 621	0.731 680	0.021 062	0.695 063
Space_ga-5	96	0.3	0.2	0.047 224	0.255 319	0.173 816	0.098 612
Space_ga-6	96	0.3	0.2	0.083 867	0.222 647	0.150 040	0.115 016
Space_ga-7	91	0.3	0.2	0.103 305	0.299 864	0.118 663	0.104 155
Space_ga-8	95	0.3	0.2	0.166 769	0.043 802	14.548 26	0.212 466
Space_ga-9	95	0.3	0.2	0.221 478	0.297 141	3.015 252	0.015 427
Space_ga-10	96	0.3	0.2	0.055 273	0.738 813	0.068 549	0.576 536

注：加粗数据为最优值；SV 为支持向量的个数； χ_{\min} 为生成粒子的尾部综合指数下界， d 为尾部综合指数的截断值，下同

表 3 本方法在 RobustLSSVM 算法上的泛化性能比较

Tab. 3 Comparison of the generalization performance of the proposed method on RobustLSSVM algorithm

数据集	SV	χ_{\min}	d	算法			
				QT-RobustLSSVM		RobustLSSVM	
				MSE	SCC	MSE	SCC
Cadata-1	95	0.3	0.2	0.267 032	0.578 401	0.280 433	0.566 52
Cadata-2	92	0.3	0.2	0.069 969	0.061 171	0.080 111	0.004 832
Cadata-3	93	0.3	0.2	0.052 688	0.090 332	0.056 108	0.106 941
Cadata-4	95	0.3	0.2	0.134 260	0.465 270	0.184 601	0.396 975
Cadata-5	97	0.3	0.2	0.039 624	0.326 844	0.039 950	0.311 920
Cadata-6	96	0.3	0.2	0.064 347	0.187 656	0.069 052	0.159 240

续表

数据集	SV	χ_{\min}	d	算法			
				QT-RobustLSSVM		RobustLSSVM	
				MSE	SCC	MSE	SCC
Cadata-7	94	0.3	0.2	0.098 245	0.594 902	0.097 20	0.478 233
Cadata-8	93	0.3	0.2	0.059 905	0.794 685	0.091 795	0.699 323
Cadata-9	93	0.3	0.2	0.061 598	0.333 843	0.089 862	0.336 031
Cadata-10	96	0.3	0.2	0.153 620	0.069 702	0.279 875	0.021 627
Cpusmall-1	97	0.3	0.2	0.005 412	0.764 837	0.006 833	0.790 982
Cpusmall-2	95	0.3	0.2	0.009 585	0.616 824	0.011 365	0.566 208
Cpusmall-3	98	0.3	0.2	0.005 623	0.796 703	0.006 806	0.763 143
Cpusmall-4	98	0.3	0.2	0.006 152	0.839 925	0.006 483	0.842 919
Cpusmall-5	98	0.3	0.2	0.004 323	0.821 184	0.004 663	0.815 045
Cpusmall-6	98	0.3	0.2	0.005 847	0.788 513	0.009 198	0.694 429
Cpusmall-7	95	0.3	0.2	0.202 723	0.428 039	0.213 005	0.365 060
Cpusmall-8	94	0.3	0.2	0.095 389	0.339 388	0.095 863	0.331 222
Cpusmall-9	97	0.3	0.2	0.031 383	0.819 635	0.031 468	0.945 307
Cpusmall-10	98	0.3	0.2	0.009 101	0.784 267	0.016 196	0.767 806
Cpusmall-11	99	0.3	0.2	0.009 693	0.867 243	0.014 528	0.787 184
Housing-1	98	0.3	0.2	0.023 595	0.838 141	0.027 799	0.805 820
Housing-2	93	0.3	0.2	0.164 104	0.631 406	0.225 335	0.534 475
Housing-3	90	0.3	0.2	0.047 909	0.809 239	0.058 919	0.743 319
Housing-4	97	0.3	0.2	0.259 533	0.368 804	0.802 485	0.310 853
Housing-5	96	0.3	0.2	0.140 546	0.142 694	0.127 464	0.164 064
Mpg-1	97	0.3	0.2	0.023 241	0.851 519	0.024 809	0.851 958
Mpg-2	93	0.3	0.2	0.058 988	0.716 925	0.068 067	0.677 861
Mpg-3	96	0.3	0.2	0.028 897	0.850 570	0.032 580	0.847 246
Mpg-4	96	0.3	0.2	0.095 770	0.480 785	0.098 530	0.506 571
Triazines-1	97	0.3	0.2	0.196 389	0.204 164	0.214 490	0.119 925
Triazines-2	97	0.3	0.2	0.220 668	0.090 920	0.214 490	0.119 925
Mg-1	91	0.5	0.2	0.286 435	0.176 377	0.305 309	0.162 251
Mg-2	96	0.5	0.2	0.042 327	0.863 890	0.033 143	0.891 027
Mg-3	96	0.5	0.2	0.141 801	0.523 961	0.159 450	0.461 358
Mg-4	91	0.5	0.2	0.145 314	0.473 626	0.160 496	0.431 148
Mg-5	95	0.5	0.2	0.078 647	0.762 762	0.072 231	0.789 298
Mg-6	90	0.5	0.2	0.083 726	0.692 422	0.085 607	0.681 127
Mg-7	91	0.5	0.2	0.049 620	0.837 254	0.050 496	0.825 314
Mg-8	89	0.5	0.2	0.068 766	0.773 502	0.098 344	0.679 806
Mg-9	90	0.5	0.2	0.106 946	0.642 502	0.110 751	0.629 421
Mg-10	93	0.5	0.2	0.066 782	0.758 531	0.071 779	0.743 342
Space_ga-1	97	0.1	0.2	0.221 118	0.139 956	0.269 337	0.201 115
Space_ga-2	96	0.1	0.2	0.657 152	0.237 010	0.842 074	0.052 882
Space_ga-3	96	0.1	0.2	0.016 411	0.897 248	0.059 016	0.842 023
Space_ga-4	97	0.1	0.2	0.018 434	0.705 332	0.016 631	0.729 877
Space_ga-5	98	0.1	0.2	0.050 805	0.238 921	0.059 374	0.186 280
Space_ga-6	94	0.1	0.2	0.081 006	0.193 692	0.092 607	0.162 074
Space_ga-7	91	0.1	0.2	0.099 215	0.359 938	0.100 637	0.299 098
Space_ga-8	94	0.1	0.2	0.124 459	0.228 870	0.155 127	0.370 834
Space_ga-9	94	0.1	0.2	0.164 141	0.428 677	0.213 760	0.418 462
Space_ga-10	96	0.1	0.2	0.048 036	0.735 042	0.048 317	0.723 019

由表2和表3可以看出,改进后的两种LSSVM的MSE值与两种传统LSSVM的MSE值相比,除一些结果相差不大外,其余均有明显减小,尤其是与标准LSSVM算法的比较结果,特别明显。另外,由表2和表3中SCC平均值的比较中发现,除一些结果相差不大外,其余均有明显提高。还需要指出的是,两种传统LSSVM算法的支持向量是全体训练样本,而改进后的两种LSSVM的支持向量个数要少一些。这主要是由于在学习训练中一些尾部样本未参与造成的。总体上说,在模型选择中引入尾部综合指数后,其预测精度与相关程度明显比未引入时更优秀。

4 结论

本文首先指出了传统的泛化误差界没有考虑训练样本预测误差的尾部分布情况。为解决这个问题,本文将这些训练样本的预测误差按其大小进行排序,然后重新讨论经验风险与结构风险,并针对无界非负损失函数集,重新推导出期望风险上界,提出了尾部综合指数,进一步分析了尾部综合指数下界对期望风险界的影响。并指出为减小结构风险,必须考虑尾部综合指数下界。为综合考虑多种泛化误差界以提升学习算法的泛化性能,本文还提出在模型选择中引入尾部综合指数下界,并将这一方法应用到LSSVM中。

[参考文献] (References)

- [1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016. (in Chinese)
- [2] CHERKASSKY V. The nature of statistical learning theory[J]. IEEE Transactions on Neural Networks, 1997, 8(6): 1564.
- [3] CHUNG K M, KAO W C, SUN C L, et al. Radius margin bounds for support vector machines with the RBF kernel[J]. Neural Computation, 2003, 15(11): 2643-2681.
- [4] VAPNIK V N. Statistical learning theory[M]. New York: John Wiley & Sons, Interscience, 1998.
- [5] VAPNIK V, CHAPELLE O. Bounds on error expectation for support vector machines[J]. Neural Computation, 2000, 12(9): 2013-2036.
- [6] GUNTER L, ZHU J. Efficient computation and model selection for the support vector regression[J]. Neural Computation, 2007, 19(6): 1633-1655.
- [7] CAWLEY G C, TALBOT N L C. Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters[J]. Journal of Machine Learning Research, 2007, 8(8): 841-861.
- [8] BARTLETT P L, MENDELSON S. Rademacher and Gaussian complexities: risk bounds and structural results[J]. The Journal of Machine Learning Research, 2003, 3: 463-482.
- [9] BARTLETT P L, BOUCHERON S, LUGOSI G. Model selection and error estimation[J]. Machine Learning, 2002, 48(1-3): 85-113.
- [10] BARTLETT P L, BOUSQUET O, MENDELSON S. Local rademacher complexities[J]. Annals of Statistics, 2005, 33(4): 1497-1537.
- [11] DEVROYE L, WAGNER T. Distribution-free performance bounds for potential function rules[J]. IEEE Transactions on Information Theory, 1979, 25(5): 601-604.
- [12] BOUSQUET O, ELISSEEFF A. Stability and generalization[J]. Journal of Machine Learning Research, 2002, 2(3): 499-526.
- [13] POGGIO T, RIFKIN R, MUKHERJEE S, et al. General conditions for predictivity in learning theory[J]. Nature, 2004, 428(6981): 419-422.
- [14] DING L, LIAO S. Model selection with the covering number of the ball of RKHS[C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York: ACM, 2014: 1159-1168.

- [15] SMALE S, ZHOU D X. Learning theory estimates via integral operators and their approximations[J]. *Constructive Approximation*, 2007, 26(2): 153-172.
- [16] GUO Z C, ZHOU D X. Concentration estimates for learning with unbounded sampling[J]. *Advances in Computational Mathematics*, 2013, 38(1): 207-223.
- [17] LV S G, FENG Y L. Integral operator approach to learning theory with unbounded sampling[J]. *Complex Analysis & Operator Theory*, 2012, 6(3): 533-548.
- [18] EBERTS M, STEINWART I. Optimal regression rates for SVMs using Gaussian kernels[J]. *Electronic Journal of Statistics*, 2013, 7(7): 1-42.
- [19] HUANG P S, AVRON H, SAINATH T N, et al. Kernel methods match deep neural networks on TIMIT[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2014: 205-209.
- [20] 冯昌, 廖士中. 随机傅里叶特征空间中高斯核支持向量机模型选择[J]. *计算机研究与发展*, 2016, 53(9): 1971-1978.
FENG C, LIAO S Z. Model selection for Gaussian kernel support vector machines in random Fourier feature space[J]. *Journal of Computer Research and Development*, 2016, 53(9): 1971-1978. (in Chinese)
- [21] 冯昌, 廖士中. 大规模核方法的随机假设空间方法[J]. *计算机科学与探索*, 2018, 12(5): 785-793.
FENG C, LIAO S Z. Large-scale kernel methods via random hypothesis spaces[J]. *Journal of Frontiers of Computer Science and Technology*, 2018, 12(5): 785-793. (in Chinese)
- [22] LIU D B, MA H, LI J H. Error distribution modeling of embedded sensors on smartphones by using laser ranger[C]//2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Hong Kong: IEEE, 2017: 387-392.
- [23] FISHER R A, TIPPETT L H C. Limiting forms of the frequency distribution of the largest or smallest member of a sample[J]. *Mathematical Proceedings of the Cambridge Philosophical Society*, 1928, 24(2): 180-190.
- [24] 刘维奇, 赫英迪, 陈琳. 重尾分布的尾部指数估计及沪深股市实证分析[J]. *数学的实践与认识*, 2011, 41(6): 1-13.
LIU W Q, HE Y D, CHEN L. Tail index estimation of heavy-tailed distribution and empirical analysis of China's stock markets[J]. *Journal of Mathematics in Practice and Theory*, 2011, 41(6): 1-13. (in Chinese)
- [25] SAHU S, SINGH H V, KUMAR B. A heavy-tailed levy distribution for despeckling ultrasound image[C]//GUPTA P, TYAGI V, GHRERA S P. Proceedings of 2017 Fourth International Conference on Image Information Processing (ICIIP). Shimla: IEEE, 2017: 1-5.
- [26] HILL B. A simple general approach to inference about the tail of a distribution[J]. *Annals of Statistics*, 1975, 3(5): 1163-1174.
- [27] 刘维奇, 邢红卫. 重尾指数估计中阈值 k 的简便优化估计[J]. *系统工程理论与实践*, 2010, 30(8): 1465-1470.
LIU W Q, XING H W. Simple and optimized choice of the threshold for heavy-tailed index estimation[J]. *Systems Engineering-Theory & Practice*, 2010, 30(8): 1465-1470. (in Chinese)
- [28] 刘勇, 廖士中. 基于支持向量机泛化误差界的多核学习方法[J]. *武汉大学学报(理学版)*, 2012, 58(2): 149-156.
LIU Y, LIAO S Z. Multiple kernel learning with the generalization error bound of support vector machine[J]. *Journal of Wuhan University (Natural Science Edition)*, 2012, 58(2): 149-156. (in Chinese)
- [29] UCHIDA M. Tight lower bound of generalization error in ensemble learning[C]//2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS). Kitakyushu: IEEE, 2015: 1130-1133.
- [30] YE Q, HUO H, ZHU T, et al. Harbor selection in large-scale remote sensing images using both deep-learned and topological structure features[C]//2017 10th International Symposium on Computational Intelligence and Design (ISCID). Hangzhou, 2017: 218-222.
- [31] CHEN L, WANG J, LI L. The models of granular system and algebraic quotient space in granular computing[J]. *Chinese Journal of Electronics*, 2016, 25(6): 1109-1113.
- [32] CHEN X, HAN Z, WANG Y, et al. A generalized model for robust tensor factorization with noise modeling by mixture of Gaussians[J]. *IEEE Transactions on Neural Networks & Learning Systems*, 2018(99): 1-14.

- [33] YONG H, MENG D, ZUO W, et al. Robust online matrix factorization for dynamic background subtraction[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 40(7): 1726-1740.
- [34] 余晓东, 雷英杰, 岳韶华, 等. 基于粒子群优化的直觉模糊核聚类算法研究[J]. 通信学报, 2015, 36(5): 78-84.
YU X D, LEI Y J, YUE S H, et al. Research on PSO-based intuitionistic fuzzy kernel clustering algorithm[J]. Journal on Communications, 2015, 36(5): 78-84. (in Chinese)
- [35] 吴彦文, 王洁. 基于混合遗传粒子群优化推荐算法的设计[J]. 计算机工程与设计, 2017, 38(2): 405-408.
WU Y W, WANG J. Similarity recommendation based on hybrid of GA and PSO[J]. Computer Engineering and Design, 2017, 38(2): 405-408. (in Chinese)
- [36] ASUNCION A, NEWMAN D J. UCI Machine learning repository irvine[DB/OL]. [2007-03-01]. <http://archive.ics.uci.edu/ml/index.php>.
- [37] SUYKENS J A K, GESTEL T V, BRABANTER J D, et al. Least squares support vector machines[J]. International Journal of Circuit Theory & Applications, 2002, 27(6): 605-615.
- [38] SUYKENS J A K, BRABANTER J D, LUKAS L, et al. Weighted least squares support vector machines: robustness and sparse approximation[J]. Neurocomputing, 2002, 48(1): 85-105.
- [39] BRABANTER K, PELCKMANS K, BRABANTER J, et al. Robustness of kernel based regression: a comparison of iterative weighting schemes[C]//International Conference on Artificial Neural Networks. Berlin: Springer-Verlag, 2009: 100-110.
- [40] BRABANTER J D, PELCKMANS K, SUYKENS J A K, et al. Robust cross-validation score function for non-linear function estimation[J]. Lecture Notes in Computer Science, 2002, 2415: 713-719.
- [41] NELDER J A, MEAD R. A simplex method for function minimization[J]. The Computer Journal, 1965, 7(4): 308-313.
- [42] XAVIER-DE-SOUZA S, SUYKENS J A K, VANDEWALLE J, et al. Coupled simulated annealing[J]. IEEE Transactions on Systems Man & Cybernetics. Part B: Cybernetics, 2010, 40(2): 320-335.