文章编号: 1001-0920 (2012) 02-0205-06

基于 Help-Training 的半监督支持向量回归

程玉虎, 冀 杰, 王雪松

(中国矿业大学 信息与电气工程学院，江苏 徐州 221116)

摘 要: 提出一种基于 Help-Training 的半监督支持向量回归算法, 包含最小二乘支持向量回归 (LS-SVR) 和 k 近邻 (kNN) 两种类型学习器. 主学习器 LS-SVR 通过选择高置信度的未标记样本加以标记, 并将其添加到已标记样本集, 使训练样本的规模不断扩大, 以提高 LS-SVR 的函数逼近性能. 辅学习器 kNN 用以协助 LS-SVR 从训练样本比较密集的区域选取未标记样本加以置信度评估, 可以减弱噪声对学习效果的负面影响. 实验结果表明所提算法具有良好的回归估计性能, 学习精度较高.

关键词: 半监督学习；助训练；支持向量回归；k 近邻；置信度

中图分类号: TP18 **文献标识码:** A

Semi-supervised support vector regression based on Help-Training

CHENG Yu-hu, JI Jie, WANG Xue-song

(School of Information and Electrical Engineering，China University of Mining and Technology，Xuzhou 221116，China. Correspondent：CHENG Yu-hu，E-mail：chengyuhu@163.com)

Abstract: A semi-supervised support vector regression based on Help-Training is proposed, which includes two kinds of learners: a least squares support vector regression(LS-SVR) and a k-nearest neighbor(kNN). As a main learner, the LS-SVR chooses unlabeled samples with the highest confidence to label and adds these samples to the labeled sample set, which is repeated for given iterations to enlarge the scale of the training samples so as to improve the property of function approximation of the LS-SVR. As an auxiliary learner, the kNN is used to help the LS-SVR choose unlabeled samples to evaluate confidence from a high-density region of training samples, which can weaken the negative influence of noise on the learning performance of the LS-SVR. Experimental results show that the Help-Training LS-SVR has advantages of good regression performance and high learning accuracy.

Key words: semi-supervised learning；Help-Training；support vector regression；k nearest neighbor；confidence

1 引 言

支持向量机 (SVM) 是一种有监督的学习方法, 需要足够的有标记样本来建立数学模型. 而在一些实际应用中, 如医疗诊断、垃圾邮件过滤等, 获得已标记的训练样本集的过程往往不仅效率低下, 而且价格昂贵, 数量有限, 因此传统 SVM 的优势难以发挥. 然而, 与此同时, 大量的未标记样本可以自动廉价地获取, 如何把未标记样本引入到学习中, 是当今机器学习的研究热点. 半监督学习的目的是将大量未标记样本与少量的已标记样本相结合以提高学习器的泛化能力[1].

1977 年, Vapnik 提出了直推式学习方法, 引起了人们极大关注, 即在推理阶段直接使用未标记样本所含有的信息, 训练出来的分类器便具有较好的泛化性能[2]. 后来, 陆续有学者将直推学习技术引入 SVM 中, 提出了一些半监督 SVM 学习方法, 典型的包括: Joachims[3]提出的直推支持向量机 TSVM, 它假定未标记样本是测试例, 通过直接推断来优化特定测试集上的性能; Bennett 等人[4]将 SVM 规范化的表现形式与半监督学习技术中的聚类假设有效地结合, 提出一种半监督支持向量机 S^3VM. 但由于 S^3VM 同时利用了有标记和未标记样本来最大化分类间隔, 导致其目标函数是非凸的. 因此, 目前 S^3VM 的研究热点主要集中在寻找有效的近似算法, 如 Chapelle 等人[5] 利用

收稿日期: 2010-08-01；**修回日期:** 2010-11-01.

基金项目: 国家自然科学基金项目(60804022, 60974050, 61072094)；教育部新世纪优秀人才支持计划项目(NCET-08-0836)；霍英东NCET-10-0765青年教师基金项目(121066)；教育部博士点基金项目(20110095110016).

作者简介: 程玉虎(1973−), 男, 教授, 博士, 从事机器学习的研究；王雪松(1974−), 女, 教授, 博士, 从事机器学习、生物信息学等研究.

梯度下降法解决非凸优化问题, 提出了梯度下降S^{3}VM; Chapelle等人[6]采用连续化技术对半监督支持向量机的非凸优化问题进行求解, 提出了cS^{3}VM; Collobert等人[7]通过凹凸过程来解决非凸优化问题, 提出了CCCPS^{3}VM.

除直推学习外, 自训练(Self-Training)和助训练(Help-Training)是半监督学习的另两种重要的学习技术, 将其与SVM相结合, 构成了Self-Training SVM和Help-Training SVM. 在Self-Training SVM学习过程中, 首先用少量有标记样本训练出一个初始SVM分类器; 然后分类器再对无标记样本进行标记; 最后选取置信度最高的标记过的无标记样本, 并将其加入训练集中重新训练更新SVM分类器, 以此提高SVM的泛化性能[8]. 但是, 初始已标记样本集规模很小, SVM分类精度较低, 在Self-Training的训练过程中, 误标记相当数量的样本是不可避免的, 这样会在已标记样本集中引入大量的噪声, 影响SVM的分类性能. Help-Training是Self-Training的一种改进形式, Adankon等人[9]通过引入Parzen窗估计作为辅学习器来对未标记样本进行预筛选, 选择属于不同类别概率较高的未标记样本作为候选样本提供给SVM分类器进行训练, 这样可以避免将误标记样本添加到已标记样本集中, 从而改进Self-Training SVM的分类性能.

尽管目前已经提出了各种类型的半监督SVM学习方法, 并已在文本分类脑机接口、生物信息学等领域得到成功应用, 但是现有研究主要用于解决SVM的半监督分类问题. 即便回归研究和分类同等重要, 但关于半监督支持向量回归的研究和利用还非常缺乏. 基于上述分析, 将半监督学习思想引入到最小二乘支持向量回归(LS-SVR)中, 提出一种基于Help-Training的半监督LS-SVR, 通过在人工和实际数据集上的对比实验, 说明了算法的有效性.

2 基于Help-Training的半监督最小二乘支持向量回归

2.1 算法流程

半监督学习的基本思想是利用数据分布上的模型假设, 建立学习器对未标记样本进行标记[1]. 半监督学习的基本设置是给定一个来自某未知分布的已标记样本集$L=\{(x_1,y_1),(x_2,y_2),\cdots,(x_{|L|},y_{|L|})\}$以及一个未标记样本集$U=\{x'_1,x'_2,\cdots,x'_{|U|}\}$, 期望学得函数$f:X\to Y$, 可以准确地对样本$x$预测其标记$y$. 此处, $x_i,x'_i\in X$; $y_i\in Y$为样本x_i的标记, 对于回归问题, 这里的y是实值输出并具有平滑性, 聚类假设一般不成立, 而流形假设仍然成立; $|L|$和$|U|$分别为L和U的大小, 即它们所包含的样本数.

图1给出了基于Help-Training的半监督LS-SVR的基本算法流程图, 图中主要包含两种类型的学习器: 主学习器LS-SVR和辅学习器k近邻(kNN). LS-SVR是标准SVR的一种扩展, 它把SVR的学习问题转化为解线性方程组问题, 因此具有较快的运算速度, 降低了SVR的学习难度[10]. 在初始学习阶段, 半监督学习可以利用的仅是少量的有标记样本, 而LS-SVR对于小样本问题有很好的学习特性. 因此, 此处选用LS-SVR作为主学习器可以获得更好的预测结果, 并且可以提高算法的运算速度. 采用kNN作为辅学习器是基于以下两点考虑: 首先, 在半监督回归学习中, 学习器需要经过很多次迭代才能变得精确, 而kNN回归模型不需要单独的训练阶段, 相比于神经网络等算法需要单独的训练阶段更加有效率; 其次, 为了选出合适的未标记样本作为候选样本, 需要对标记的置信度进行估计, 这需要考虑回归估计中流形假设的局部平滑性, 而kNN可以通过利用邻近训练样本的属性来很好地满足这种局部平滑性.

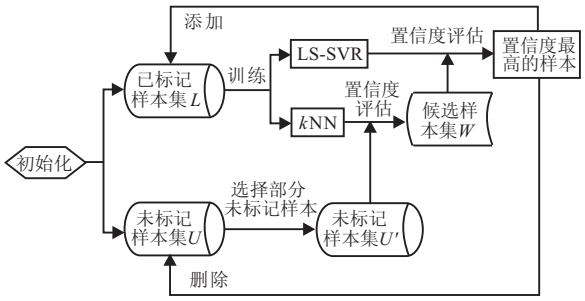


图1 算法流程图

在Help-Training LS-SVR的学习过程中, 首先利用已标记样本集L来训练LS-SVR和kNN; 其次, 从未标记样本集U中选择部分未标记样本组成新的未标记样本集U', kNN通过对U'中的样本进行置信度评估, 选择出具有最高置信度的未标记样本组成候选样本集W; 然后, LS-SVR通过对候选样本集W中的样本进行置信度评估, 选择置信度最高的样本添加到标记样本集L中, 同时在未标记样本集U中删除相应的样本. 重复这个过程, 可以增加已标记样本集的数量, 提高LS-SVR学习的精度.

2.2 置信度评估

在半监督回归学习中, 样本的置信度评估是一个非常重要的问题, 它直接关系到算法是否对未标记样本进行正确的标记, 并影响算法的性能. 此处需要分别考虑kNN和LS-SVR两种学习器在半监督回归中的置信度评估.

为了能够通过kNN选出合适的候选样本, 需要对样本标记置信度进行评估, 找到置信度最高的样本. 在分类问题中, 样本标记置信度的评估相对简单, 可

以通过比较未标记样本属于不同类别的概率来评估. 但是, 回归问题的类别标签是连续的实值, 因此很难找到这样的估计概率. 此处借鉴 COREG 算法中的评估机制: 如果利用具有最高置信度的样本, 则在标记样本集上训练的学习器的回归误差减小的最多, 也就是说, 置信度最高的样本是指把样本加入标记样本集中使标记样本的预测值与实际实值标签值最一致的样本[11].

在利用 kNN 进行候选样本集构造时, 首先, 对于 U' 中每一个未标记样本 x_u, 从已标记样本集 L 中找出 x_u 的 k 个近邻组成近邻样本集 Ω_u; 其次, 计算利用 (x_u, y_u) 之后得到的新 kNN 学习器 h' 在 Ω_u 上的均方误差, 其中 $x_u \in U'$ 为未标记样本, y_u 为初始 kNN 学习器 h 对 x_u 的实值预测标签, $y_u = h(x_u)$; 然后, 计算初始学习器 h 与新学习器 h' 在 Ω_u 上的均方误差之间的差值 Δ_u, 每一个未标记样本对应一个 Δ_u, 与最大 Δ_u 值对应的 x_u 被认为是置信度最高的样本; 最后, 将置信度最高的样本加入候选样本集 W 中. 置信度最高的未标记样本可通过将下式最大化来评估:

$$\Delta_u = \sqrt{\sum_{x_i \in \Omega_u} (y_i - h(x_i))^2 \Big/ k} - \sqrt{\sum_{x_i \in \Omega_u} (y_i - h'(x_i))^2 \Big/ k}. \tag{1}$$

其中: $h(x_i)$ 表示 h 对 x_i 的回归估计值; $h'(x_i)$ 表示 h' 对 x_i 的回归估计值; y_i 表示 x_i 的实际实值标签值; k 为近邻数, 一般可以通过距离度量 D 来选择未标记样本的 k 个近邻标记样本, D 可以是欧氏距离、马氏距离、闵氏距离等.

LS-SVR 作为主学习器需要通过对候选样本集中的样本进行置信度评估, 找到置信度最高的样本加入到标记样本集中. 与 kNN 的评估机制相似, 首先, 利用标记样本集 L 训练 LS-SVR, 得到初始学习器 g, 预测候选样本集中的样本 x_j, 得到其预测值 y_j, 即 $y_j = g(x_j)$. 把候选样本集 W 中的样本 (x_j, y_j) 加入到标记样本集 L 中得到新的标记样本集 L'; 然后利用 L' 训练 LS-SVR, 得到新的学习器 g', 计算 g' 在标记样本集 L 上的均方误差 Λ_i, 与最小 Λ_i 值所对应的样本 (x_j, y_j) 即为置信度最高的样本, 即通过将下式最小化来评估:

$$\Lambda_i = \sqrt{\sum_{i=1}^{|L|} (y_i - y_i')^2 \Big/ |L|}. \tag{2}$$

其中: (x_i, y_i) 为 L 中的标记样本, y_i 为 x_i 的实值标签; y_i' 为 g' 对 x_i 的预测值, 即 $y_i' = g'(x_i)$.

2.3 最小二乘支持向量回归

假设有训练样本集 $L = \{(x_1, y_1), (x_2, y_2), \cdots, (x_{|L|}, y_{|L|})\}$, 则有回归函数为

$$y(x) = \omega\phi(x) + b. \tag{3}$$

其中: $\phi(x)$ 为特征映射, ω 为权值向量, b 为偏置项.

LS-SVR 算法的目标是求解如下最小值问题[12]:

$$\begin{cases} \min J(\omega, \xi) = \dfrac{1}{2}\|\omega\| + \dfrac{\gamma}{2}\sum\limits_{i=1}^{|L|}\xi_i^2, \\ \text{s.t. } y_i = \omega\phi(x_i) + b + \xi_i. \end{cases} \tag{4}$$

其中: γ 为正则化参数, ξ_i 为误差的松弛变量. 构造 Lagrange 函数

$$L(\omega, b, \xi, \alpha) = \frac{1}{2}\|\omega\| + \frac{\gamma}{2}\sum_{i=1}^{|L|}\xi_i^2 - \sum_{i=1}^{|L|}a_i(\omega\phi(x_i) + b + \xi_i - y_i), \tag{5}$$

其中 $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_{|L|}]^{\mathrm{T}}$ 为 Lagrange 乘子. 根据 Karush-Kuhn-Tucker 条件, 对 ω, b, ξ_i, α_i 求偏导数, 并令其为零, 优化问题可以变成求解线性方程

$$\begin{bmatrix} 0 & eI^{\mathrm{T}} \\ eI & F + \gamma^{-1}I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ \hat{Y} \end{bmatrix}. \tag{6}$$

其中: $\hat{Y} = [y_1, y_2, \cdots, y_{|L|}]^{\mathrm{T}}$, $eI = [1, \cdots, 1]^{\mathrm{T}}$, $F = K(x_i, x_{i'}) = \phi(x_i)^{\mathrm{T}} \cdot \phi(x_{i'})$ 为满足 Mercer 条件的核函数.

通过求解式 (6), 可得最小二乘支持向量机的回归模型为

$$y(x) = \sum_{i=1}^{|L|}\alpha_i K(x_i, x_{i'}) + b. \tag{7}$$

3 算法步骤

基于 Help-Training 的半监督最小二乘支持向量回归算法步骤如下:

Step 1: 初始化参数. 包括已标记样本集 L, 未标记样本集 U, 最大迭代次数 T; LS-SVR 的核函数类型 $K(x_i, x_{i'})$, 正则化参数 γ 和核宽度参数 σ^2; kNN 的距离度量 D 和 k 值.

Step 2: 把已标记样本集 L 作为训练样本集, 从未标记样本集 U 中随机选取 n 个样本组成新的未标记样本集 U'. 利用标记样本集 L 训练 LS-SVR 得到初始学习器 g, 训练 kNN 得到初始学习器 h.

Step 3: 对于 U' 中的每一个未标记样本 x_u, 计算其近邻样本集 Ω_u. 把 (x_u, y_u) 加入 L 后训练 kNN 得到新学习器 h'. 根据式 (1), 得到 n 个未标记样本对应的 Δ_u 值.

Step 4: 从 U' 中选择与最大 Δ_u 值所对应的未标记样本 x_u 作为置信度最高的样本, 将其放入候选样本集 W 中.

Step 5: 对于 W 中的每一个样本 x_j, 利用 g 预测其实值标签 y_j. 把 (x_j, y_j) 加入 L 后训练 LS-SVR 得到

新学习器 g'. 根据式 (2), 得到 W 中的每一个样本 x_j 对应的 Λ_i 值, 选择最小 Λ_i 值对应的样本 (x_j, y_j) 为置信度最高的样本.

Step 6: 把标记置信度最高的样本 (x_j, y_j) 加入标记样本集 L 中, 并从未标记样本集 U 中删除 x_j. 返回 Step 2, 重复训练 T 代.

Step 7: 利用加入未标记样本后的标记样本集 L 训练 LS-SVR, 并输出 LS-SVR 的回归参数 α 和 b, 得到回归模型 (7).

4 实验及结果分析

为了验证算法的有效性, 在 9 个人工模拟数据集和 2 个 UCI 实际数据集上进行了实验, 关于数据集的描述分别如表 1 和表 2 所示, 其中 $U[a_1, a_2]$ 表示自变量的取值在区间 $[a_1, a_2]$ 上均匀分布, 样本个数为样本总个数, 包括训练样本和测试样本.

分别将监督 LS-SVR, 半监督 Self-Training LS-SVR 和半监督 Help-Training LS-SVR 用于解决上述实验数据集的回归估计问题. 在实验过程中, 选择数据总数的 75% 作为训练样本, 25% 作为测试样本. 训练样本中分别选择 10% 和 30% 作为已标记样本, 其余的样本去掉其实值标签作为未标记样本. 对于所有数据集的数据, 将其输入变量和实值标签归一化处理为 [0, 1] 区间. 另外, 对于人工数据集, 对每一组数据人为加入均值为 0, 方差为 0.01 的高斯白噪声. 所有类型 LS-SVR 的核函数均取为径向基核函数, 通过交叉验证法来选择正则化参数 γ 和核宽度 σ^2. 对半监督 LS-SVR 而言, 每一代选择 $n = 200$ 个未标记样本进行训练, 共进行 $T = 100$ 代训练; Help-Training LS-SVR 中 kNN 学习器中的距离度量 D 选择欧氏距离, $k = 3$. 采用测试样本的均方误差衡量 LS-SVR 的回归估计性能, 定义为

$$\mathrm{MSE} = \sqrt{\sum_{l=1}^{N} \varepsilon_l^2 \Big/ N}. \tag{8}$$

其中: ε_l 为第 l 个测试样本的逼近误差, N 为测试样本

表 2 实际数据集

数据集	属性值	数据个数
Concrete_Data	9	1 030
Housing_Data	14	506

表 1 人工数据集

数据集	函数表达式	自变量取值范围	样本个数
3-d Mexican Hat	$y = \mathrm{sinc}\sqrt{x_1^2 + x_2^2} = \dfrac{\sin\sqrt{x_1^2 + x_2^2}}{\sqrt{x_1^2 + x_2^2}}$	$x_1, x_2 \sim U[-4\pi, 4\pi]$	1 000
Friedman #1	$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$	$x_1, x_2, x_3, x_4, x_5 \sim U[0, 1]$	500
Friedman #2	$y = \sqrt{x_1^2 + \left(x_2 x_3 - \left(\dfrac{1}{x_2 x_4}\right)\right)^2}$	$x_1 \sim U[0, 100],\ x_2 \sim U[40\pi, 560\pi]$ $x_3 \sim U[0, 1],\ x_4 \sim U[1, 11]$	500
Friedman #3	$y = \tan^{-1}\dfrac{x_2 x_3 - \dfrac{1}{x_2 x_4}}{x_1}$	$x_1 \sim U[0, 100],\ x_2 \sim U[40\pi, 560\pi]$ $x_3 \sim U[0, 1],\ x_4 \sim U[1, 11]$	500
Gabor	$y = \dfrac{\pi}{2}\exp\left[-2(x_1^2 + x_2^2)\right]\cos\left[2\pi(x_1 + x_2)\right]$	$x_1, x_2 \sim U[0, 1]$	1 000
Multi	$y = 0.79 + 1.27x_1x_2 + 1.56x_1x_4 + 3.42x_2x_5 + 2.06x_3x_4x_5$	$x_1, x_2, x_3, x_4, x_5 \sim U[0, 1]$	1 000
Plane	$y = 0.6x_1 + 0.3x_2$	$x_1, x_2 \sim U[0, 1]$	500
Polynomial	$y = 1 + 2x + 3x^2 + 4x^3 + 5x^4$	$x \sim U[0, 1]$	500
SinC	$y = \dfrac{\sin x}{x}$	$x \sim U[0, 2\pi]$	500

表 3 测试样本的平均均方误差

数据集	30%标记样本(MSE)			10%标记样本(MSE)		
	监督LS-SVR	Self-Training LS-SVR	Help-Training LS-SVR	监督LS-SVR	Self-Training LS-SVR	Help-Training LS-SVR
3-d Mexican Hat	0.045 3	0.033 3	**0.032 2**	0.061 3	0.059 6	**0.056 5**
Friedman #1	0.047 9	0.034 5	**0.029 8**	0.110 7	0.113 1	**0.106 8**
Friedman #2	0.021 4	0.019 6	**0.014 7**	0.023 1	0.021 2	**0.018 9**
Friedman #3	1.96E-3	2.14E-3	**1.59E-3**	3.65E-3	3.82E-3	**2.98E-3**
Gabor	0.111 7	0.058 1	**0.051 4**	0.185 0	0.151 7	**0.122 4**
Multi	0.196 3	0.095 8	**0.078 9**	0.241 5	0.229 5	**0.174 9**
Plane	0.025 8	0.020 6	**0.019 0**	0.190 1	0.078 0	**0.055 4**
Polynomial	0.310 4	0.157 9	**0.051 5**	0.065 9	0.060 2	**0.044 4**
SinC	0.082 7	0.061 5	**0.047 9**	0.204 0	0.227 5	**0.143 9**
Concrete_Data	0.013 0	0.009 8	**0.009 1**	0.014 5	0.011 1	**0.009 4**
Housing_Data	0.013 2	0.008 9	**0.007 9**	0.013 4	0.011 9	**0.010 6**

个数. 对每组数据集重复 20 次实验, 取 MSE 的平均值作为最后输出. 表 3 给出了 3 种类型 LS-SVR 在 9 个人工数据集和 2 个实际数据集上, 分别选择 30% 和 10% 标记样本进行实验的平均均方误差结果.

由表 3 可以看出:

1) 对于表 1 中大多数实验数据集 (Polynomial 除外)而言, 不论是监督 LS-SVR, 还是半监督 LS-SVR, 均拥有较多的标记样本 (30% 标记率) 使学习器的学习更充分, 预测效果更好, 这完全符合 LS-SVR 有导师示教的学习规律.

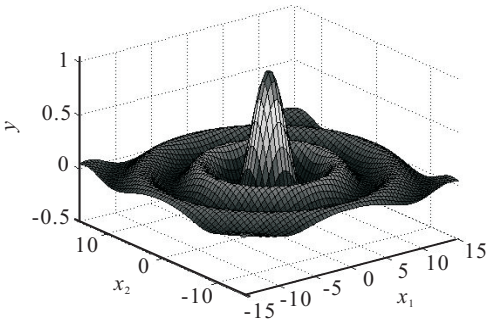
2) 对于表 1 中大多数实验数据集而言, 半监督 LS-SVR 的学习效果优于监督 LS-SVR. 这是因为, 在学习过程中, 半监督学习方式不但利用了有标记样本对单个样本精确描述的优势, 而且发挥了无标记样本对样本集整体描述的重要作用, 从而使训练出的 LS-SVR 具有更好的泛化性能, 充分体现了半监督学习的优势.

3) 一般而言, 半监督学习优于传统的监督学习方式. 但是, 对于 Friedman#1, Friedman #3 和 SinC 数据集而言, 半监督 Self-Training LS-SVR 的学习精度反而低于监督 LS-SVR. 这是由于一方面支持向量机本身具有小样本学习特性, 当标记样本量较少时, 监督 LS-SVR已经能够获得较好的学习性能. 另一方面, 由于置信度高的未标记样本有可能逼近误差较大, 在添加标记样本的同时引入了大量噪声, 从而导致 Self-Training LS-SVR 的函数逼近精度下降.

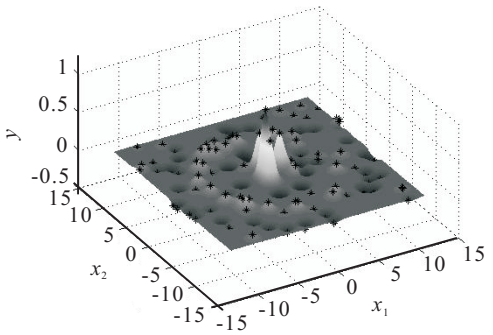
4) 在 3 种类型最小二乘支持向量回归算法中, 不论是 10% 还是 30% 的样本标记率, 半监督 Help-Training LS-SVR 在所有数据集上的平均均方误差最小, 学习效果最好. 这是由于利用了 kNN 来协助选择高置信度的未标记样本加以标记, 能够避免 Self-Training 技术选择大逼近误差的未标记样本作为高置信度样本添加到训练样本中, 减弱了噪声对学习效果的负面影响.

5) 对于半监督学习而言, 增加已标记样本的数量一般会提高学习器的学习性能. 但是, 当标记样本增加到一定程度时, 半监督算法的学习精度将不再有改进, 反而会有所降低, 如数据集 Polynomial. 这可能与模型的假设不符合真实的情况有关. 另外, 随着训练不断进行, 自动标记样本的噪声会不断积累, 其负作用会越来越大.

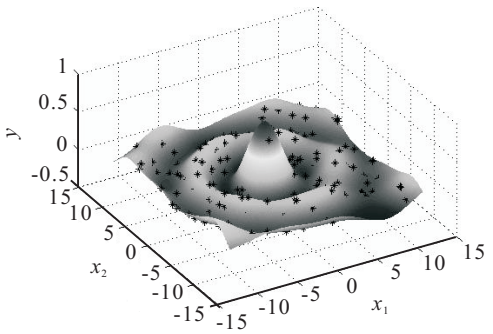
为更直观地显示最小二乘支持向量回归模型的函数逼近效果, 图 2 给出了 3 种 LS-SVR 在函数 3-d Mexican Hat 上的实验结果. 其中图 2(a) 为原函数的实际图形, 图 2(b)～图 2(d) 上的黑点表示训练样本. 由于在学习过程中, 半监督 LS-SVR 要迭代地将未标记样本加以标记并添加到标记样本集中, 因此图 2(c)～图 2(d) 上的训练样本个数要比图 2(b) 上的多, 从而使得半监督 LS-SVR 的回归精度也有很大的提高, 说明半监督学习利用未标记样本来帮助训练可以很好地改善学习性能. 在半监督 Help-Training LS-SVR 的学习过程中, 不同于 Self-Training 直接选择高置信度未标记样本进行标记的工作方式, LS-SVR 通过对候选样本集中的样本进行置信度评估来选择未标记样本



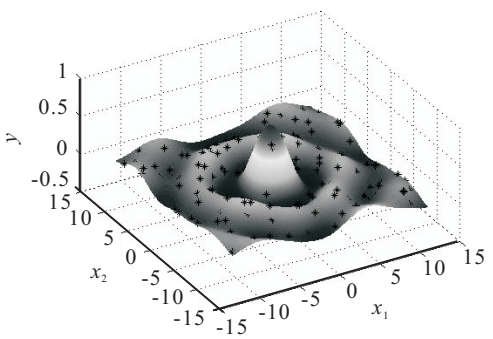
(a) 原函数图形



(b) 监督LS-SVR(MSE=0.091 4)



(c) 半监督Self-Training LS-SVR(MSE=0.059 1)

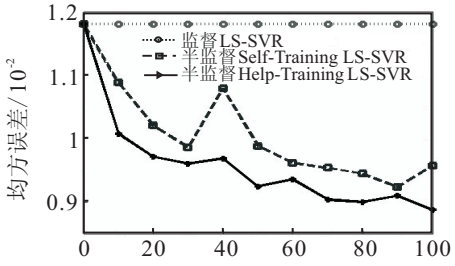


(d) 半监督Helf-Training LS-SVR(MSE=0.049 2)

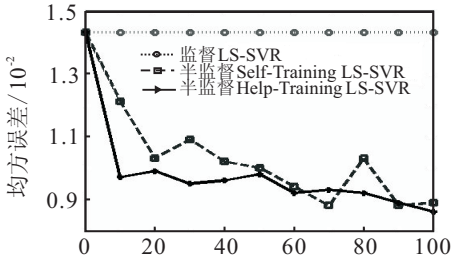
图 2　3-d Mexican Hat 实验结果对比

加以标记, 而候选样本集中的样本是利用kNN学习器从训练样本比较密集的区域进行扩展, 因此, Help-Training LS-SVR可以很好地满足流形假设的局部平滑性, 有效提高LS-SVR的学习精度.

图3和图4为两个实际数据集某次实验的测试样本均方误差与迭代次数之间的关系曲线. 为便于图形显示, 仅记录下每隔10代的均方误差值, 加之实验中设置迭代次数为$T = 100$, 因此, 每幅图中仅给出了11个数据点(包括初始状态). 监督LS-SVR的训练样本集是事先给定的, 不会随迭代次数的增加而变化, 因此, 监督LS-SVR的均方误差在整个迭代过程中固定不变, 在图中表现为一条水平直线. 对于半监督LS-SVR而言, 已标记样本的数量是逐代增加的, 因此, LS-SVR模型的学习精度得以不断改进, 测试样本的均方误差随迭代次数的增加呈减小趋势. 由图3和图4还可以看出, 半监督LS-SVR在两个实际数据集上的均方误差均小于监督LS-SVR, 而Help-Training LS-SVR的均方误差最小, 具有最好的估计性能, 这一点也与表3中的结论相符.

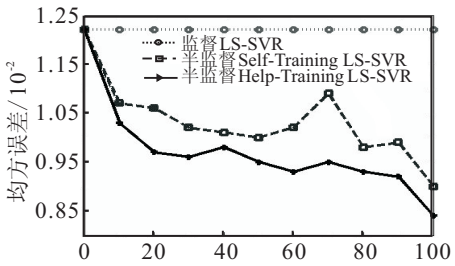


(a) 30%标记样本

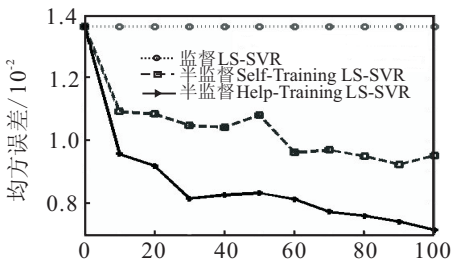


(b) 10%标记样本

图 3 测试样本均方误差变化曲线
(Concrete_Data数据集)



(a) 30%标记样本



(b) 10%标记样本

图 4 测试样本均方误差变化曲线
(Housing_Data数据集)

5 结 论

本文针对小部分已标记样本和大部分无标记样本的回归学习问题, 提出一种基于Help-Training的半监督支持向量回归算法. 首先, 利用已标记样本集训练主学习器LS-SVR和辅学习器kNN; 其次, kNN通过对未标记样本集中的样本进行置信度评估, 选择出具有最高置信度的未标记样本组成候选样本集; 然后, LS-SVR通过对候选样本集中的样本进行置信度评估, 选择置信度最高的样本添加到标记样本集中, 同时在未标记样本集中删除相应的样本. 利用更新后的已标记样本集重新训练LS-SVR, 以此提高LS-SVR的学习精度. 此处kNN用以协助LS-SVR从训练样本比较密集的区域选取未标记样本加以标记, 并添加到已标记样本集中, 可以很好地满足流形假设的局部平滑性. 人工模拟和UCI实际数据集上的实验结果验证了Help-Training LS-SVR具有良好的回归估计性能.

参考文献(References)

[1] Chapelle O, Schölkopf B, Zien A. Semi-supervised learning[M]. London: MIT Press, 2006.

[2] Vapnik V, Sterin A. On structural risk minimization or overall risk in a problem of pattern recognition[J]. Automation and Remote Control, 1977, 10(3): 1495-1503.

[3] Joachims T. Transductive inference for text classification using support vector machines[C]. Proc of the 16th Int Conf on Machine Learning. USA: Association for Computing Machinery, 1999: 200-209.

[4] Bennett K P, Demiriz A. Semi-supervised support vector machine[C]. Proc of Advances in Neural Information Processing Systems. USA: The MIT Press, 1998: 368-374.

[5] Chapelle O, Zien A. Semi-supervised classification by low density separation[C]. Proc of the 10th Int Workshop on Artificial Intelligence and Statistics. USA: Morgan Kaufmann Publishers Inc, 2005: 57-64.

[6] Chapelle O, Chi M, Zien A. A continuation method for semi-supervised SVMs[C]. Proc of the 23rd Int Conf on Machine Learning. USA: Association for Computing Machinery, 2006: 185-192.

(下转第226页)