

考虑梯度信息的 ε -支持向量回归机

周晓剑¹

摘 要 传统的 ε -支持向量回归机 (ε -support vector regression, ε -SVR) 只是根据样本点处的响应值来构建模型, 并没考虑样本点处的梯度信息. 如果样本点处的梯度信息容易获得或者获得的成本并不高, 那就应该将梯度信息应用到模型的构建中. 已有的基于梯度信息的 ε -支持向量回归机模型的构建是从泰勒展开的角度着手, 简单地将梯度信息插入到泰勒展开式中; 本研究另辟蹊径, 并没有去估计样本点邻域内的函数值, 而是将梯度信息作为第二类变量融入到核矩阵中直接构建优化模型, 使模型的构建更为简捷直观, 并据此得到一种新的基于梯度信息的 ε -支持向量回归机 (Gradient-enhanced ε -support vector regression, GESVR) 模型. 所提模型通过了常用分析函数及精算领域中的生命表数据的验证, 实验表明, 与传统的 ε -SVR 相比, 考虑梯度信息的 GESVR 模型显著地提高了其预测精度.

关键词 ε -支持向量回归机, 元模型, 梯度信息, 计算机实验设计, 仿真优化

引用格式 周晓剑. 考虑梯度信息的 ε -支持向量回归机. 自动化学报, 2014, 40(12): 2908–2915

DOI 10.3724/SP.J.1004.2014.02908

Enhancing ε -support Vector Regression with Gradient Information

ZHOU Xiao-Jian¹

Abstract Traditional methods constructing of ε -support vector regression (ε -SVR) do not consider the gradients of the true function but only deal with the exact responses at the samples. If the gradient information is available easily and cheaply, it should be used to enhance the model. The existing research on constructing of ε -SVR with gradient information starts from the perspective of Taylor expansion, and simply inserts the additional objective values in the neighborhood of the sampled points into the corresponding terms of a Taylor expansion. In this paper, the gradient-enhanced ε -support vector regression (GESVR) is developed with a direct formulation by incorporating the gradient information into the kernel matrix. The efficiency of this technique is verified by analytical function fitting and the actuarial data in the life table. The results show that the GESVR provides more reliable prediction results than ε -SVR alone.

Key words ε -support vector regression (ε -SVR), metamodel, gradient information, computer design, simulation optimization

Citation Zhou Xiao-Jian. Enhancing ε -support vector regression with gradient information. *Acta Automatica Sinica*, 2014, 40(12): 2908–2915

支持向量机 (Support vector machine, SVM) 最早可追溯到 20 世纪 70 年代, 并在 20 世纪末由 AT & T 贝尔实验室的 Vapnic 及其同事将 SVM 进一步发展; 随后由 Smola 等^[1] 将其用于回归, 即支持向量回归机 (Support vector regression, SVR). 有关 SVM 的算法的研究近年来为众多学者所关注^[2–4]. 传统的 SVR 有一些很好的特性, 如解的稀疏性, 并克服了神经网络的局部极值和过学习

的问题, 但它没有考虑到样本点处的梯度信息, 这使得其在小样本情况下拟合精度并不理想. 要想提高模型的精度, 就不得不通过采集更多的样本点并计算对应的函数值来得到更精确的模型. 在很多情形下, 计算对应样本点的函数值需要耗费财力、物力及大量的时间. 基于此, 周晓剑等^[5] 探讨了有关基于梯度信息的模型的构建, 提出了一种基于梯度信息的 ε -支持向量回归机 (ε -support vector regression, ε -SVR) 的算法, 在这一算法中, 作者采用的是“间接 ε -SVR”的方法. 该方法通过一阶泰勒展开的方式并考虑了样本点 (x_i, y_i) 以及其邻近点 $(x_i + \Delta x_i, y_i + d_i \cdot \Delta x_i)$ 和 $(x_i - \Delta x_i, y_i - d_i \cdot \Delta x_i)$ 三点处的误差, 其中 d_i 和 Δx_i 分别表示第 i 个样本点处的导数和邻域的半径. 事实上, “间接 ε -SVR”的方法是通过泰勒展开增加了样本点的个数, 通过增加样本点来提高模型的预测能力, 这种做法使得

收稿日期 2013-12-26 录用日期 2014-07-18

Manuscript received December 26, 2013; accepted July 18, 2014
国家自然科学基金 (71401080), 江苏省高校哲学社会科学基金项目 (2013SJB6300072), 南京邮电大学人才引进项目 (NYS212008) 资助
Supported by National Natural Science Foundation of China (71401080), University Social Science Fund of Jiangsu (2013SJB6300072), and Talent Introduction Foundation of Nanjing University of Posts and Telecommunications (NYS212008)

本文责任编辑 刘成林

Recommended by Associate Editor LIU Cheng-Lin

1. 南京邮电大学管理学院 南京 210023

1. School of management, Nanjing University of Posts and Telecommunications, Nanjing 210023

模型的构建更为复杂, 而且需要人为地去设定泰勒展开式中参数 Δx_i 的值, 这在实际当中有时并不好控制. 本文在此另辟蹊径, 提出一种“直接 ε -SVR”的方法, 该方法并不是从增加样本点的角度出发, 模型的构建并不需要参数 Δx_i , 因此并不需要人为地设定参数 Δx_i 的值, 而是直接将梯度信息植入 ε -SVR 的核函数矩阵中, 通过求解核函数的偏导数来提高模型的预测能力.

1 ε -支持向量回归机 (ε -SVR)

给定数据集 $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)\}$ (其中 l 代表样本数目), 若 SVR 的损失函数取 ε 不敏感函数

$$L_\varepsilon(f(\mathbf{x}) - y) = \begin{cases} 0, & \text{若 } |f(\mathbf{x}) - y| < \varepsilon \\ |f(\mathbf{x}) - y| - \varepsilon, & \text{其他} \end{cases} \quad (1)$$

则构成 ε -支持向量回归机 (ε -SVR), 本文中称 ε 为 ε -SVR 的参数, 其模型如下:

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \quad (2)$$

$$\text{s.t.} \begin{cases} y_i - (\mathbf{w} \cdot \mathbf{x}_i + b) \leq \varepsilon + \xi_i \\ (\mathbf{w} \cdot \mathbf{x}_i + b) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, \quad i = 1, \dots, l \end{cases}$$

其中, ξ_i, ξ_i^* 为松弛变量, C 为惩罚参数.

该模型的解可由以下二次规划问题求得:

$$\min \frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^l (\alpha_i - \alpha_i^*) y_i + \varepsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) \quad (3)$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0, \\ 0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \dots, l \end{cases}$$

其中, $K(\cdot, \cdot)$ 是核函数. 在求出上式中的参数 α, α^* 后, 即可求得回归函数 $f(\mathbf{x})$. 使用 ε -不敏感函数并利用矩阵方法求解式 (3) 时, 矩阵具有稀疏性.

以上模型只考虑了样本点处的函数值, 并未考虑样本点处的梯度信息. 当梯度信息容易得到时, 则可以考虑构建基于梯度信息的 ε -支持向量回归机 (Gradient-enhanced ε -support vector regression, GESVR) 模型, 下面将给出具体的构建过程.

2 基于梯度信息的 ε -支持向量回归机 (GESVR)

2.1 原始最优化问题

对于给定的样本集, 将输入变量 X 映射到特征空间: $F = \{\phi(\mathbf{x}) | \mathbf{x} \in X\}$, SVR 的基本形式可表示为

$$f(\mathbf{x}) = \mathbf{w} \phi(\mathbf{x}) + b \quad (4)$$

则 $f(\mathbf{x})$ 的一阶偏导数为

$$D^r(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_r} = \mathbf{w} \frac{\partial \phi(\mathbf{x})}{\partial x_r}, \quad r = 1, \dots, d \quad (5)$$

其中, x_r 是 $\mathbf{x}$ 的第 r 维.

如果没有梯度信息可用, 我们将求解如下二次规划问题:

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C_0 \sum_{i=1}^l (\xi_i^* + \xi_i) \quad (6)$$

$$\text{s.t.} \begin{cases} y_i - \mathbf{w} \phi(\mathbf{x}_i) - b \leq \varepsilon_0 + \xi_i \\ \mathbf{w} \phi(\mathbf{x}_i) + b - y_i \leq \varepsilon_0 + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, \quad i = 1, \dots, l \end{cases}$$

否则如果梯度信息容易获得, 与传统的 ε -SVR 一样, 对于梯度信息我们仍然需要寻找到一个“好的”超平面, 使其能够平分最近点. 因此, 针对梯度信息可以得到相应的最优化问题:

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^l \sum_{r=1}^d C_r (\zeta_{ir}^* + \zeta_{ir}) \quad (7)$$

$$\text{s.t.} \begin{cases} D^r(\mathbf{x}_i) - \mathbf{w} \cdot \frac{\partial \phi(\mathbf{x}_i)}{\partial \mathbf{x}_{ir}} \leq \varepsilon_r + \zeta_{ir} \\ \mathbf{w} \cdot \frac{\partial \phi(\mathbf{x}_i)}{\partial \mathbf{x}_{ir}} - D^r(\mathbf{x}_i) \leq \varepsilon_r + \zeta_{ir}^* \\ \zeta_{ir}, \zeta_{ir}^* \geq 0, \quad i = 1, \dots, l; \quad r = 1, \dots, d \end{cases}$$

注意到最优化问题 (6) 和 (7) 分别是针对样本数据和梯度信息而设计的. 为了将梯度信息融入到 ε -SVR 模型的构建当中, 我们将问题 (6) 和 (7) 融合得到一个新的最优化问题:

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C_0 \sum_{i=1}^l (\xi_i + \xi_i^*) + \sum_{i=1}^l \sum_{r=1}^d C_r (\zeta_{ir} + \zeta_{ir}^*) \quad (8)$$

$$\text{s. t. } \begin{cases} y_i - \mathbf{w}\phi(\mathbf{x}_i) - b \leq \varepsilon_0 + \xi_i \\ D^r(\mathbf{x}_i) - \mathbf{w} \frac{\partial \phi(\mathbf{x}_i)}{\partial \mathbf{x}_{ir}} \leq \varepsilon_r + \zeta_{ir} \\ \mathbf{w}\phi(\mathbf{x}_i) + b - y_i \leq \varepsilon_0 + \xi_i^* \\ \mathbf{w} \frac{\partial \phi(\mathbf{x}_i)}{\partial \mathbf{x}_{ir}} - D^r(\mathbf{x}_i) \leq \varepsilon_r + \zeta_{ir}^* \\ \xi_i^{(*)}, \zeta_{ir}^{(*)} \geq 0, \quad i = 1, \dots, l; \quad r = 1, \dots, d \end{cases}$$

其中, $(*)$ 包括了带星号的以及不带星号的向量, 如 $\xi_i^{(*)} \geq 0$ 表示 $\xi_i \geq 0$ 及 $\xi_i^* \geq 0$, 我们用 $\xi^{(*)}$ 表示 $(\xi_1^*, \xi_1, \dots, \xi_l^*, \xi_l)^T$; 类似地, $\zeta_{ir}^{(*)} \geq 0$ 意味着 $\zeta_{ir} \geq 0$ 以及 $\zeta_{ir}^* \geq 0$ ($i = 1, \dots, l; r = 1, \dots, d$); 我们用 $\zeta^{(*)}$ 来代表 $(\zeta_1, \dots, \zeta_d)^T$, 其中 $\zeta_r = (\zeta_{1r}^*, \zeta_{1r}, \dots, \zeta_{lr}^*, \zeta_{lr})^T, r = 1, \dots, d$.

一方面, 参数 C_0 (或 C_r ($r = 1, \dots, d$)) 用于权衡函数 f (或者 D^r ($r = 1, \dots, d$)) 的平坦性及超出超平面中 ε_0 (或者 ε_r ($r = 1, \dots, d$)) “管道” 的样本的离差和; 另一方面, C_0 (或 C_r ($r = 1, \dots, d$)) 还可以权衡目标函数中样本信息及梯度信息的重要性. 当 $C_r = 0$ ($r = 1, \dots, d$) 时, GESVR 就等价于传统的 SVR, 这可以通过后面将给出的式 (10) 看出, 因为 $C_r = 0$ 意味着 $\beta_{ir}^{(*)} = 0$, 这说明传统的 SVR 是 GESVR 的一个特例, 通过选择合适的参数我们可以得到比传统 SVR 更精确的模型, 其优越性将在第 3 节给出说明.

通过求解原始最优化问题 (8), 我们可以得到最优解 $(\hat{\mathbf{w}}, \hat{b}, \hat{\xi}^{(*)}, \hat{\zeta}^{(*)})$, 进而得到决策函数:

$$\hat{f}(\mathbf{x}) = \hat{\mathbf{w}}^T \cdot \mathbf{x} + \hat{b} \quad (9)$$

注意到 $(\hat{\mathbf{w}}, \hat{b}, \hat{\xi}^{(*)}, \hat{\zeta}^{(*)})$ 中的 $(\hat{\xi}^{(*)}, \hat{\zeta}^{(*)})$ 在此并不起作用, 我们只需关注 $(\hat{\mathbf{w}}, \hat{b})$, 因此有如下定义:

定义 1. 称 $(\hat{\mathbf{w}}, \hat{b})$ 为原始最优化问题 (8) 关于 $(\mathbf{w}, b)$ 的解, 如果存在 $\xi^{(*)} \in \mathbf{R}^{2l}$ 及 $\zeta^{(*)} \in \mathbf{R}^{2ld}$ 使得 $(\hat{\mathbf{w}}, \hat{b}, \hat{\xi}^{(*)}, \hat{\zeta}^{(*)})$ 是原始最优化问题 (8) 的解. 类似地, 称 $\hat{\mathbf{w}}$ 为原始最优化问题 (8) 关于 $\mathbf{w}$ 的解, 如果存在 $\hat{b} \in \mathbf{R}, \xi^{(*)} \in \mathbf{R}^{2l}$ 及 $\zeta^{(*)} \in \mathbf{R}^{2ld}$ 使得 $(\hat{\mathbf{w}}, \hat{b}, \hat{\xi}^{(*)}, \hat{\zeta}^{(*)})$ 是原始最优化问题 (8) 的解.

定理 1. 原始最优化问题 (8) 关于 $(\mathbf{w}, b)$ 的解存在.

定理 2. 原始最优化问题 (8) 关于 $\mathbf{w}$ 的解存在且唯一的.

但原始最优化问题 (8) 关于 b 的解却并不一定是唯一的, 事实上, 当 ε_r ($r = 0, 1, \dots, d$) 足够大时, 原始最优化问题 (8) 就有形如 $(\bar{\mathbf{w}}, \bar{b}, \bar{\xi}^{(*)}, \bar{\zeta}^{(*)}) = (\mathbf{0}, \bar{b}, \mathbf{0}, \mathbf{0})$ 的解, 其中 $\bar{b}$ 有无穷多的取值可能.

2.2 对偶问题及其与原始问题的关系

定理 3. 原始最优化问题 (8) 的对偶问题为

$$\begin{aligned} \min_{\substack{\alpha^{(*)} \in \mathbf{R}^{2l} \\ \beta^{(*)} \in \mathbf{R}^{2ld}}} & \frac{1}{2} \left[\sum_{i=1}^l (\alpha_i^* - \alpha_i) \phi(\mathbf{x}_i) + \sum_{i=1}^l \sum_{r=1}^d (\beta_{ir}^* - \beta_{ir}) \frac{\partial \phi(\mathbf{x}_i)}{\partial \mathbf{x}_{ir}} \right] \times \\ & \left[\sum_{j=1}^l (\alpha_j^* - \alpha_j) \phi(\mathbf{x}_j) + \sum_{j=1}^l \sum_{r=1}^d (\beta_{jr}^* - \beta_{jr}) \frac{\partial \phi(\mathbf{x}_j)}{\partial \mathbf{x}_{jr}} \right] + \\ & \varepsilon_0 \sum_{i=1}^l (\alpha_i^* + \alpha_i) + \\ & \varepsilon_r \sum_{i=1}^l \sum_{r=1}^d (\beta_{ir}^* + \beta_{ir}) - \sum_{i=1}^l y_i (\alpha_i^* - \alpha_i) - \\ & \sum_{i=1}^l \sum_{r=1}^d D^r(\mathbf{x}_i) (\beta_{ir}^* - \beta_{ir}) \end{aligned} \quad (10)$$

$$\text{s. t. } \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0,$$

$$0 \leq \alpha_i^{(*)} \leq C_0, \quad i = 1, \dots, l,$$

$$0 \leq \beta_{ir}^{(*)} \leq C_r, \quad i = 1, \dots, l; \quad r = 1, \dots, d$$

为便于问题的推导, 我们在此将对偶问题 (10) 写成矩阵的形式. 则对偶问题 (10) 关于 $\alpha^{(*)}$ 和 $\beta^{(*)}$ 求极大值等价于:

$$\min_{\gamma \in \mathbf{R}^{2(d+1)l}} W(\gamma) = \frac{1}{2} \gamma^T \psi \gamma - \tilde{\mathbf{y}}^T \gamma \quad (11)$$

$$\text{s. t. } \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0,$$

$$0 \leq \alpha_i^{(*)} \leq C_0, \quad i = 1, \dots, l$$

$$0 \leq \beta_{ir}^{(*)} \leq C_r, \quad i = 1, \dots, l; \quad r = 1, \dots, d$$

其中

$$\gamma = (\gamma_0^T, \gamma_1^T, \dots, \gamma_d^T)^T$$

$$\gamma_0 = \alpha^{(*)} = (\alpha_1^*, \alpha_1, \dots, \alpha_l^*, \alpha_l)^T$$

$$\gamma_r = \beta_r = (\beta_{1r}^*, \beta_{1r}, \dots, \beta_{lr}^*, \beta_{lr})^T, \quad r = 1, \dots, d$$

$$\tilde{\mathbf{y}} = [\tilde{\varepsilon} + \dot{\mathbf{y}}^T, \tilde{\varepsilon} - \dot{\mathbf{y}}^T]^T$$

$$\dot{\mathbf{y}} = [\mathbf{y}^T, (D^1)^T, \dots, (D^d)^T]^T$$

$$D^r = (D^r(\mathbf{x}_1), D^r(\mathbf{x}_2), \dots, D^r(\mathbf{x}_l))^T$$

$$\tilde{\varepsilon} = [\mathbf{e}\varepsilon_0, \mathbf{e}\varepsilon_1, \dots, \mathbf{e}\varepsilon_d]$$

$$\mathbf{e} = \mathbf{1}_{1 \times l}$$

式 (11) 中的核矩阵 ψ 可表示为

$$\psi = \begin{bmatrix} \dot{\psi} & -\dot{\psi} \\ -\dot{\psi} & \dot{\psi} \end{bmatrix} \quad (12)$$

其中

$$\dot{\psi} = \begin{bmatrix} H_{00} & H_{11} & \cdots & H_{1d} \\ H_{11} & Q_{11} & \cdots & Q_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ H_{1d} & Q_{d1} & \cdots & Q_{dd} \end{bmatrix} \quad (13)$$

$$H_{00} = K(\mathbf{x}_i, \mathbf{x}_j) =$$

$$\begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_l) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_l, \mathbf{x}_1) & \cdots & k(\mathbf{x}_l, \mathbf{x}_l) \end{bmatrix} \quad (14)$$

$$H_{1r} = \frac{\partial K(\mathbf{x}_i, \mathbf{x}_j)}{\partial x_{ir}} =$$

$$\begin{bmatrix} \frac{\partial k(\mathbf{x}_1, \mathbf{x}_1)}{\partial x_{1r}} & \cdots & \frac{\partial k(\mathbf{x}_1, \mathbf{x}_l)}{\partial x_{lr}} \\ \vdots & \ddots & \vdots \\ \frac{\partial k(\mathbf{x}_l, \mathbf{x}_1)}{\partial x_{1r}} & \cdots & \frac{\partial k(\mathbf{x}_l, \mathbf{x}_l)}{\partial x_{lr}} \end{bmatrix} \quad (15)$$

$$H_{r1} = \frac{\partial K(\mathbf{x}_i, \mathbf{x}_j)}{\partial x_{jr}} =$$

$$\begin{bmatrix} \frac{\partial k(\mathbf{x}_1, \mathbf{x}_1)}{\partial x_{1r}} & \cdots & \frac{\partial k(\mathbf{x}_1, \mathbf{x}_l)}{\partial x_{lr}} \\ \vdots & \ddots & \vdots \\ \frac{\partial k(\mathbf{x}_l, \mathbf{x}_1)}{\partial x_{1r}} & \cdots & \frac{\partial k(\mathbf{x}_l, \mathbf{x}_l)}{\partial x_{lr}} \end{bmatrix} \quad (16)$$

$$Q_{rm} = \frac{\partial K^2(\mathbf{x}_i, \mathbf{x}_j)}{\partial x_{ir} \partial x_{jm}} =$$

$$\begin{bmatrix} \frac{\partial k^2(\mathbf{x}_1, \mathbf{x}_1)}{\partial x_{1r} \partial x_{1m}} & \cdots & \frac{\partial k^2(\mathbf{x}_1, \mathbf{x}_l)}{\partial x_{1r} \partial x_{lm}} \\ \vdots & \ddots & \vdots \\ \frac{\partial k^2(\mathbf{x}_l, \mathbf{x}_1)}{\partial x_{lr} \partial x_{1m}} & \cdots & \frac{\partial k^2(\mathbf{x}_l, \mathbf{x}_l)}{\partial x_{lr} \partial x_{lm}} \end{bmatrix} \quad (17)$$

在式 (15) ~ (17) 中, $r, m = 1, \dots, d$, 且核函数为高斯核, 这一核函数在机器学习领域广为采用, 其

相应的偏导数如下:

$$\frac{\partial k(\mathbf{x}_i, \mathbf{x}_j)}{\partial x_{ir}} = -\frac{1}{\delta^2}(x_{ir} - x_{jr})k(\mathbf{x}_i, \mathbf{x}_j) \quad (18)$$

$$\frac{\partial k(\mathbf{x}_i, \mathbf{x}_j)}{\partial x_{jr}} = \frac{1}{\delta^2}(x_{ir} - x_{jr})k(\mathbf{x}_i, \mathbf{x}_j) \quad (19)$$

$$\frac{\partial k^2(\mathbf{x}_i, \mathbf{x}_j)}{\partial x_{ir} \partial x_{jm}} =$$

$$\begin{cases} \left[\frac{1}{\delta^2} - \frac{1}{\delta^4}(x_{ir} - x_{jr})^2 \right] k(\mathbf{x}_i, \mathbf{x}_j), & r = m \\ -\frac{1}{\delta^4}(x_{ir} - x_{jr})(x_{im} - x_{jm})k(\mathbf{x}_i, \mathbf{x}_j), & r \neq m \end{cases} \quad (20)$$

注意到最优化问题 (11) 并不一定是严格的凸规划问题, 因此其解并不唯一. 以下我们将从对偶问题 (10) 的任一解出发求出原始问题 (8) 关于 $(\mathbf{w}, b)$ 的解.

定理 4. 设 $\bar{\gamma} = (\bar{\gamma}_0^T, \bar{\gamma}_1^T, \dots, \bar{\gamma}_d^T)^T$ (其中 $\bar{\gamma}_0 = (\bar{\alpha}_1^*, \bar{\alpha}_1, \dots, \bar{\alpha}_l^*, \bar{\alpha}_l)^T$, $\bar{\gamma}_r = (\bar{\beta}_{1r}^*, \bar{\beta}_{1r}, \dots, \bar{\beta}_{lr}^*, \bar{\beta}_{lr})^T, r = 1, \dots, d$) 是对偶问题 (10) 的一个解, 则原始问题 (8) 关于 $\mathbf{w}$ 的唯一解可表示为

$$\bar{\mathbf{w}} = \sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) \phi(\mathbf{x}_i) + \sum_{i=1}^l \sum_{r=1}^d (\bar{\beta}_{ir}^* - \bar{\beta}_{ir}) \frac{\partial \phi(\mathbf{x}_i)}{\partial x_{ir}} \quad (21)$$

在定理 4 中, 我们已说明乘子 $\bar{b}$ 是原始问题关于 b 的解, 接下来的定理 5 将给出 $\bar{b}$ 的计算方法.

定理 5. 设 $\bar{\gamma} = (\bar{\gamma}_0^T, \bar{\gamma}_1^T, \dots, \bar{\gamma}_d^T)^T$ (其中 $\bar{\gamma}_0 = (\bar{\alpha}_1^*, \bar{\alpha}_1, \dots, \bar{\alpha}_l^*, \bar{\alpha}_l)^T$, $\bar{\gamma}_r = (\bar{\beta}_{1r}^*, \bar{\beta}_{1r}, \dots, \bar{\beta}_{lr}^*, \bar{\beta}_{lr})^T, r = 1, \dots, d$) 是对偶问题 (10) 的解, 定义

$$S_1^0 = \{i | \bar{\alpha}_i = 0\}, S_1 = \{i | \bar{\alpha}_i \in (0, C)\}$$

$$S_1^C = \{i | \bar{\alpha}_i = C\} \quad (22)$$

$$S_2^0 = \{i | \bar{\alpha}_i^* = 0\}, S_2 = \{i | \bar{\alpha}_i^* \in (0, C)\}$$

$$S_2^C = \{i | \bar{\alpha}_i^* = C\} \quad (23)$$

并对 $j = \{1, 2\}, k = \{0, C\}$ 定义

$$d_{\max}(S_j^k) = \max_{\mathbf{x} \in S_j^k} \{y_i - \bar{\mathbf{w}}\phi(\mathbf{x}_i)\} \quad (24)$$

$$d_{\min}(S_j^k) = \min_{\mathbf{x} \in S_j^k} \{y_i - \bar{\mathbf{w}}\phi(\mathbf{x}_i)\} \quad (25)$$

则原始问题 (8) 关于 $(\mathbf{w}, b)$ 的解集可表示为

$$\{(\mathbf{w}, b) | \mathbf{w} = \bar{\mathbf{w}}, b = \bar{b} \in [b_{dn}, b^{up}]\} \quad (26)$$

其中

$$\bar{\mathbf{w}} = \sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) \phi(\mathbf{x}_i) + \sum_{i=1}^l \sum_{r=1}^d (\bar{\beta}_{ir}^* - \bar{\beta}_{ir}) \frac{\partial \phi(\mathbf{x}_i)}{\partial x_{ir}}$$

并且 b_{dn} 和 b^{up} 可按下列相应情形计算:

1) 如果 $S_1 \neq \emptyset$, 则对任意的 $j \in S_1$, 按如下方式计算 b_{dn} 和 b^{up} :

$$b_{dn} = b^{up} = y_j - \left[\sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^l \sum_{r=1}^d (\bar{\beta}_{ir}^* - \bar{\beta}_{ir}) \frac{\partial k(\mathbf{x}_i, \mathbf{x}_j)}{\partial x_{ir}} \right] - \varepsilon_0 \quad (27)$$

2) 如果 $S_2 \neq \emptyset$, 则对任意的 $k \in S_2$, 按如下方式计算 b_{dn} 和 b^{up} :

$$b_{dn} = b^{up} = y_k - \left[\sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) k(\mathbf{x}_i, \mathbf{x}_k) + \sum_{i=1}^l \sum_{r=1}^d (\bar{\beta}_{ir}^* - \bar{\beta}_{ir}) \frac{\partial k(\mathbf{x}_i, \mathbf{x}_k)}{\partial x_{ir}} \right] + \varepsilon_0 \quad (28)$$

3) 如果 $S_1 = \emptyset$ 以及 $S_2 = \emptyset$, 则按如下方式计算 b_{dn} 和 b^{up} :

$$b^{up} = \min\{d_{\max}(S_2^C) - \varepsilon_0, d_{\max}(S_1^0) + \varepsilon_0\} \quad (29)$$

$$b_{dn} = \min\{d_{\max}(S_1^C) + \varepsilon_0, d_{\max}(S_2^0) - \varepsilon_0\} \quad (30)$$

2.3 构建基于梯度信息的 ε -支持向量回归机 (GESVR)

本节建立基于梯度信息的 ε -支持向量回归机 (GESVR) 模型. 显然, 如果我们得到了对偶问题 (10) 的解 $\bar{\gamma} = (\bar{\gamma}_0^T, \bar{\gamma}_1^T, \dots, \bar{\gamma}_d^T)^T$ (其中 $\bar{\gamma}_0 = (\bar{\alpha}_1^*, \bar{\alpha}_1, \dots, \bar{\alpha}_l^*, \bar{\alpha}_l)^T$, $\bar{\gamma}_r = (\bar{\beta}_{1r}^*, \bar{\beta}_{1r}, \dots, \bar{\beta}_{lr}^*, \bar{\beta}_{lr})^T, r = 1, \dots, d$), 根据定理 4 和定理 5 我们可以得到原始问题所有的解. 这里关于 $\bar{b}$ 的计算应该包括定理 5 的三种情形, 但在实际应用中往往只发生情形 1 和情形 2, 因此我们重点针对以上两种情形建立 GESVR 模型.

GESVR 模型的构建算法:

步骤 1. 给定训练集 $T = \{(\mathbf{x}_1, y_1, D_{11}, \dots, D_{1d}), \dots, (\mathbf{x}_l, y_l, D_{l1}, \dots, D_{ld})\}$, 其中 $\mathbf{x}_i \in \mathbf{R}^n; y_i, D_{ir} \in \mathbf{R}, i = 1, \dots, l, r = 1, \dots, d$;

步骤 2. 选择适当的参数 ε_r 和 C_r ($r = 0, 1, \dots, d$), 并选择合适的核函数 (本文选用高斯核函数);

步骤 3. 构建并求解最优化问题

$$\min_{\gamma \in \mathbf{R}^{2(d+1)l}} W(\gamma) = \frac{1}{2} \gamma^T \Psi \gamma - \tilde{\mathbf{y}}^T \gamma \quad (31)$$

$$\text{s. t. } \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0$$

$$0 \leq \alpha_i^{(*)} \leq C_0, \quad i = 1, \dots, l$$

$$0 \leq \beta_{ir}^{(*)} \leq C_r, \quad i = 1, \dots, l; r = 1, \dots, d$$

得到最优解 $\bar{\gamma} = (\bar{\gamma}_0^T, \bar{\gamma}_1^T, \dots, \bar{\gamma}_d^T)^T$ (其中 $\bar{\gamma}_0 = (\bar{\alpha}_1^*, \bar{\alpha}_1, \dots, \bar{\alpha}_l^*, \bar{\alpha}_l)^T$, $\bar{\gamma}_r = (\bar{\beta}_{1r}^*, \bar{\beta}_{1r}, \dots, \bar{\beta}_{lr}^*, \bar{\beta}_{lr})^T, r = 1, \dots, d$);

步骤 4. 构建决策函数:

$$f(\mathbf{x}) = \sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^l \sum_{r=1}^d (\bar{\beta}_{ir}^* - \bar{\beta}_{ir}) \frac{\partial k(\mathbf{x}_i, \mathbf{x}_j)}{\partial x_{ir}} + \bar{b} \quad (32)$$

其中, $\bar{b}$ 选择下列方式计算: 选择位于开区间的 $\bar{\alpha}_j \in (0, C)$ 或者 $\bar{\alpha}_k^* \in (0, C)$. 如果选择的是 $\bar{\alpha}_j$, 则计算

$$\bar{b} = y_j - \left[\sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^l \sum_{r=1}^d (\bar{\beta}_{ir}^* - \bar{\beta}_{ir}) \frac{\partial k(\mathbf{x}_i, \mathbf{x}_j)}{\partial x_{ir}} \right] - \varepsilon_0 \quad (33)$$

否则, 如果选择的是 $\bar{\alpha}_k^*$, 则计算

$$\bar{b} = y_k - \left[\sum_{i=1}^l (\bar{\alpha}_i^* - \bar{\alpha}_i) k(\mathbf{x}_i, \mathbf{x}_k) + \sum_{i=1}^l \sum_{r=1}^d (\bar{\beta}_{ir}^* - \bar{\beta}_{ir}) \frac{\partial k(\mathbf{x}_i, \mathbf{x}_k)}{\partial x_{ir}} \right] + \varepsilon_0 \quad (34)$$

3 模型的验证与分析

本节将采用一维和二维的分析测试函数及寿险精算中的生命表数据来验证本文所提算法的有效性.

3.1 一维函数的拟合

首先, 采用 1 个经典的一维测试函数检验 GESVR 模型的有效性. 在该实验中, 等间隔地产生 6 个样本点, 样本点处的梯度信息可以解析地求得. 该测试函数定义如下^[6]:

$$f(x) = \frac{\sin(x)}{x}, \quad x \in [-10, 10] \quad (35)$$

参数 C_r, ε_r ($r = 0, 1$) 和 σ 的值采用交叉验证法加以确定, 其结果见表 1.

表 1 GESVR 和 SVR 的参数值

Table 1 The results of cross-validation for parameters in GESVR and SVR

模型	模型中参数
GESVR	$C_0 = 1, C_1 = 1.2, \varepsilon_0 = 0.001, \varepsilon_1 = 0.01, \sigma = 1.5$
SVR	$C_0 = 1, \varepsilon_0 = 0.001, \sigma = 1.5$

SVR 和 GESVR 的回归结果见图 1. 虽然 SVR 和 GESVR 在所有的 6 个样本点处都有较好的回归效果, 但 SVR 模型并不能有效地捕捉到这些函数在样本点外的真实趋势, 而 GESVR 模型不仅能捕捉到样本点处的局部信息而且还能较准确地捕捉到样本点以外的函数趋势, 这说明了 GESVR 的有效性.

3.2 二维函数的拟合

对于二维函数的拟合, 采用 CamelBack 函数来测试所提模型的有效性, 该函数为^[7]

CamelBack:

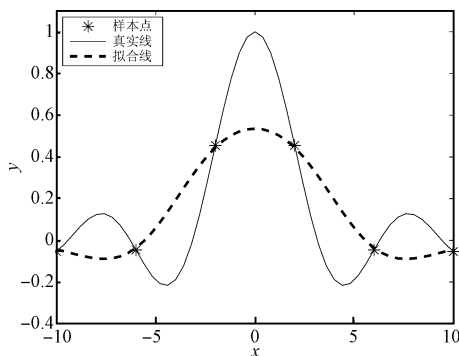
$$f(x_1, x_2) = \left(4 - 2.1x_1^2 + \frac{x_1^4}{3}\right)x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2 \quad (36)$$

其中, $x_1 \in [-3, 3], x_2 \in [-2, 2]$.

常用的评价模型精度的指标有: 均方根误差 (Root mean squared error, RMSE)、平均绝对误差 (Average absolute error, AAE)、最大绝对误差 (Max absolute error, MAE) 及相关系数 (Correlation coefficient, 用 R 表示), 它们的定义如下:

均方根误差:

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n_{\text{error}}} \frac{(y_i - \hat{y})^2}{n_{\text{error}}}} \quad (37)$$



(a) 用 SVR 回归
(a) Fitted with SVR

平均绝对误差:

$$\text{AAE} = \sum_{i=1}^{n_{\text{error}}} \frac{|y_i - \hat{y}|}{n_{\text{error}}} \quad (38)$$

最大绝对误差:

$$\text{MAE} = \max |y_i - \hat{y}|, \quad i = 1, \dots, n_{\text{error}} \quad (39)$$

相关系数:

$$R(y, \hat{y}) = \frac{\frac{1}{V} \int_V (y - \bar{y})(\hat{y} - \bar{\hat{y}}) dv}{\delta(y)\delta(\hat{y})} \quad (40)$$

$$\frac{1}{V} \int_V y \hat{y} dv = \sum_{i=1}^{n_{\text{error}}} \frac{y_i \hat{y}_i}{n_{\text{error}}}$$

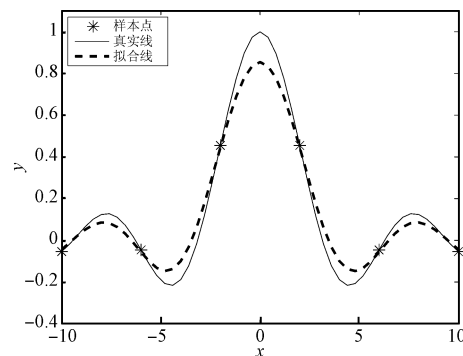
$$\bar{y} = \sum_{i=1}^{n_{\text{error}}} \frac{y_i}{n_{\text{error}}}$$

$$\delta(y) = \sqrt{\sum_{i=1}^{n_{\text{error}}} \frac{(y_i - \bar{y})^2}{n_{\text{error}}}}$$

在以上定义中, n_{error} 代表测试集的样本数, y_i 是真实的响应, $\bar{y}$ 是真实响应的平均值, $\hat{y}$ 是预测响应, $\bar{\hat{y}}$ 是预测响应的平均值. RMSE、AAE 及 R 用于评价模型整体上的精度, 而 MAE 用于衡量模型局部的精度.

对于这两个测试函数, 本实验采用拉丁超立方抽样方法进行采样. 为了考察不同样本量对结果的影响, 从 5 到 50 采取不同的样本量来构建模型, 然后抽取 35 个样本来验证它们的预测精度. 为了减少随机因素的影响, 我们重复 500 次, 并取 R , RMAE, AAE 和 MAE 的均值来衡量模型的精度.

在 CamelBack 测试中, 对于 GESVR 和 SVR 中共有的参数 C_0, ε_0 和 σ , 采用交叉验证法来确定



(b) 用 GESVR 回归
(b) Fitted with GESVR

图 1 SVR 和 GESVR 对 $f(x)$ 的回归结果

Fig. 1 Comparison of SVR and GESVR for $f(x)$

它们的值, 其值为: $C_0 = 100, \varepsilon_0 = 0.01, \sigma = 3.5$. 对于 GESVR 中特有的参数 C_r, ε_r ($r = 1, 2$), 我们发现, 当 $\varepsilon_r > 0.1$ ($r = 1, 2$) 时, GESVR 没有明显的提高, 当 $\varepsilon_r < 0.1$ ($r = 1, 2$) 时, 通过原始问题 (8) 中的不等式约束 ε_r ($r = 1, 2$) 并不能对优化问题产生明显的影响, 因此将参数设为 $\varepsilon_r = 0.1$ ($r = 1, 2$). 以上做法是将 ε_1 和 ε_2 设为同一值, 当然也可以为其设不同的值, 这就意味着将为输入变量的每一维都设置不同的“管道”. 与参数 ε_r 一样, 为了简便起见, 我们将 C_1 和 C_2 设为同一值, 因为它们都是针对一阶偏导数的惩罚因子. 本实验的重要任务之一就是针对不同的样本数目设定相应的参数值, 这样我们可以寻找参数值随样本数目变化的规律, 这同时也是有效应用 GESVR 模型的前提. 为了寻找到最优的参数值, 我们同样使用交叉验证的方法来确定它们. 参数值的变化随样本量改变而变化的情况可见表 2. 从表 2 可以看到, 参数值随着样本的增加而减少, 换句话说, 当样本量较小时需要一个较大的惩罚因子以增加梯度信息对式 (8) 中目标函数的影响, 反之亦然.

对于 CamelBack, 图 2 直观地给出了 SVR 和 GESVR 的误差比较结果, 同时相应的数值比较结果见表 3. 从这些图和表可以看出: 1) 无论哪一种模型, 其误差 (MAE、AAE 和 RMSE) 总是随样本量的增加而减少, 而 R 则随样本量的增加而增加, 因为 R 是代表相关程度的, 其值越大说明相关性越好, 也说明预测的效果越好; 2) 无论样本量如何变化, GESVR 的回归性能都要优于 SVR, 这说明加入梯度信息能有效地提高模型的预测精度; 3) 小样本情况下使用 GESVR 模型比大样本情况下使用效果更明显, 也就是说, 当样本量小时, 使用 GESVR 来提高模型的预测能力是非常有意义的.

4 结论

本文将传统的 SVR 模型与梯度信息相结合, 提出了一种新的回归模型 — GESVR, 该模型的有效性通过了几个常用分析函数的验证. 同时, 我们还将这一模型应用于保险精算领域. 实验结果表明, 提出的模型极大地提高了预测精度. 在仔细分析模型构建过程及实验结果的基础上, 我们有如下发现:

表 2 参数值的变化随样本量的变化情况

Table 2 The variation of the value of the parameters with the number of samples

样本数	CamelBack 中的 C_r ($r = 1, 2$)
5	$C_1 = C_2 = 65$
10	$C_1 = C_2 = 30$
15	$C_1 = C_2 = 25$
20	$C_1 = C_2 = 16$
25	$C_1 = C_2 = 10$
30	$C_1 = C_2 = 10$
40	$C_1 = C_2 = 10$
50	$C_1 = C_2 = 10$

1) GESVR 的回归性能总是优于 SVR, 这是因为我们采用了“直接 SVR”的方法, 即直接将梯度信息融合到核矩阵 (13) 中, 通过求核函数的偏导来得到 GESVR 模型.

2) 使用 GESVR 模型能够得到理想的回归效果, 特别是在小样本的情形下, 这意味着: 当样本的获得成本很大时, 如单件小批量的产品 (大飞机及其他的复杂产品等), 使用基于梯度信息的建模手段来节省抽样的花费将是一个很好的选择.

表 3 函数 CamelBack 中 GESVR 及 SVR 的误差比较结果

Table 3 Numerical comparison of errors between SVR and the GESVR for CamelBack

样本数目 数目	MAE		AAE		RMSE		R	
	GESVR	SVR	GESVR	SVR	GESVR	SVR	GESVR	SVR
5	79.0462	84.9797	18.0618	20.4176	25.4284	28.0127	0.5764	0.3816
10	77.4389	81.2804	14.3209	15.9673	22.5662	24.4942	0.7087	0.6348
15	75.4660	77.4835	14.1913	14.9686	22.0756	23.0960	0.7367	0.7140
20	73.4641	77.1249	14.1403	14.9379	21.7162	23.0349	0.7670	0.7402
25	73.7531	76.4368	13.7243	14.8394	21.4214	22.6838	0.7751	0.7547
30	72.9075	73.6592	13.8056	14.0566	21.3134	21.7223	0.7854	0.7662
40	71.0279	73.7079	13.5415	13.6939	20.6778	21.3251	0.7940	0.7834
50	68.9557	71.2705	13.7579	14.0321	20.4657	21.1925	0.7939	0.7936

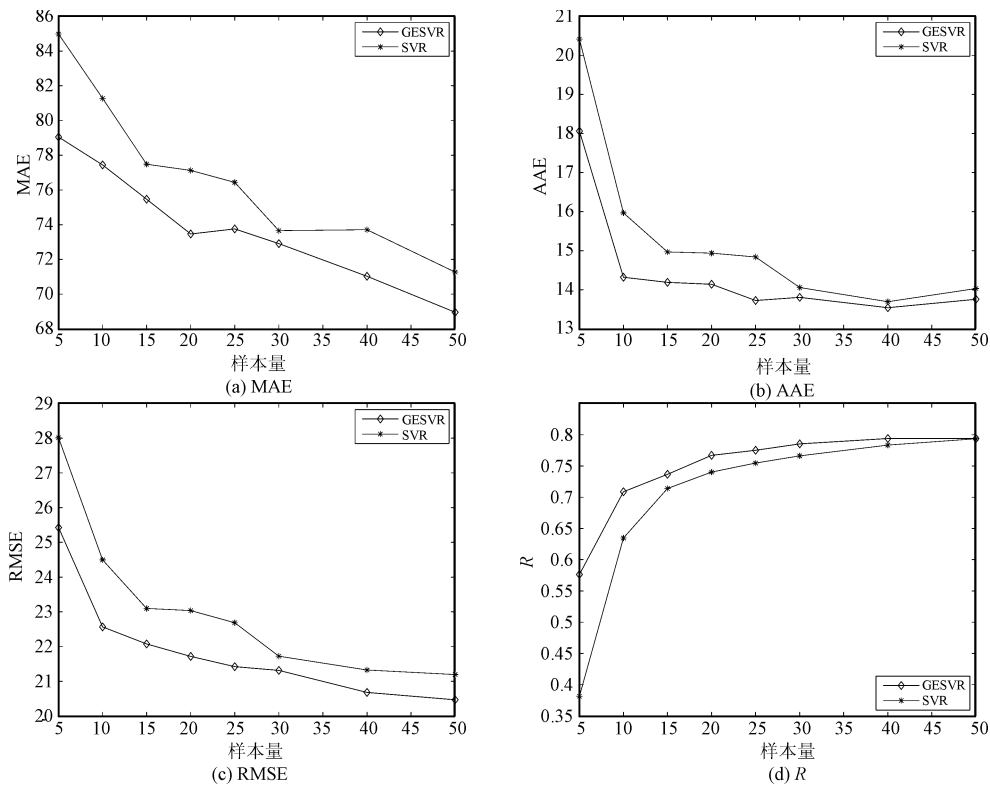


图 2 函数 CamelBack 中 GESVR 与 SVR 的误差比较结果

Fig. 2 Comparison of errors between SVR and GESVR for CamelBack

3) 参数 C_r ($r = 1, \dots, d$) (其中 d 代表样本的维度) 在模型的构建中起到至关重要的作用, 如果参数选择失当, 则模型的优势荡然无存. 实验表明: 一般而言, 样本量越小其值越大, 这是因为需要让梯度信息在建模的过程中施加更多的影响, 这体现在优化模型的目标函数 (8) 中由梯度信息导致的偏差将施以更大的惩罚因子.

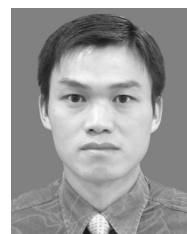
References

- 1 Smola A J, Schölkopf B. A tutorial on support vector regression. *Statistics and Computing*, 2004, **14**(3): 199–222
- 2 Tao Jian-Wen, Wang Shi-Tong. Kernel distribution consistency based local domain adaptation learning. *Acta Automatica Sinica*, 2013, **39**(8): 1295–1309
(陶剑文, 王士同. 核分布一致局部领域适应学习. 自动化学报, 2013, **39**(8): 1295–1309)
- 3 Liang Yan-Ming, Su Fang, Li Qi, Liu Ding. A self-organizing algorithm for T-S fuzzy model based on support vector machine regression and its application. *Acta Automatica Sinica*, 2013, **39**(12): 2143–2149
(梁炎明, 苏芳, 李琦, 刘丁. 基于支持向量机回归的 T-S 模糊模型自组织算法及应用. 自动化学报, 2013, **39**(12): 2143–2149)
- 4 Yang Zhi-Min, Shao Yuan-Hai, Liang Jing. Unascertained support vector machine. *Acta Automatica Sinica*, 2013, **39**(6): 895–901

39(6): 895–901

(杨志民, 邵元海, 梁静. 未确知支持向量机. 自动化学报, 2013, **39**(6): 895–901)

- 5 Zhou Xiao-Jian, Ma Yi-Zhong, Zhu Jia-Gang. Gradient-enhance support vector regression. *Systems Engineering*, 2010, **28**(3): 87–92
(周晓剑, 马义中, 朱嘉钢. 基于梯度信息的支持向量回归机. 系统工程, 2010, **28**(3): 87–92)
- 6 Alexander I, Forrester J, Keane A J. Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences*, 2009, **45**(1–3): 50–79
- 7 Zhou X J, Ma Y Z, Li X F. Ensemble of surrogates with recursive arithmetic average. *Structural and Multidisciplinary Optimization*, 2011, **44**(5): 651–671



周晓剑 南京邮电大学管理学院讲师. 主要研究方向为智能质量控制, 人工智能, 模式识别.

E-mail: xjzhou@njupt.edu.cn

(ZHOU Xiao-Jian Lecturer at Nanjing University of Posts and Telecommunications. His research interest covers intelligent quality control, artificial intelligence, and pattern recognition.)