文章编号: 1001-0920 (2015) 04-0721-06 DOI: 10.13195/j.kzyjc.2014.0044

基于自适应边界向量提取的多尺度 v-支持向量机建模

苏成利, 郑博元, 李 平

(辽宁石油化工大学 信息与控制工程学院，辽宁 抚顺 113001)

摘 要: 针对 v-支持向量机 (v-SVM) 用于大规模、多峰样本建模时易出现训练速度慢和回归精度低的问题, 提出基于边界向量提取的多尺度 v-SVM 建模方法. 该方法采用一种自适应边界向量提取算法, 从训练样本中预提取出包含全部支持向量的边界向量集, 以缩减训练样本规模, 并通过求解多尺度 v-SVM 二次规划问题获取全局最优回归模型, 从多个尺度上对复杂分布样本进行逼近. 仿真结果表明, 基于边界向量提取的多尺度 v-SVM 比 v-SVM 具有更好的回归结果.

关键词: 大样本建模；边界向量提取；多尺度学习；v-支持向量机

中图分类号: TP181 **文献标志码:** A

Multiscale v-support vector machine modeling based on adaptive boundary vector extraction

SU Cheng-li, ZHENG Bo-yuan, LI Ping

(School of Information and Control Engineering，Liaoning Shihua University，Fushun 113001，China．Correspondent: SU Cheng-li，E-mail：suchengli@lnpu.edu.cn)

Abstract: A multiscale v-support vector machine(v-SVM) based on adaptive boundary vector extraction is presented. It overcomes the disadvantages of the slow training speed and the low regression accuracy which are caused by using the general v-SVM for large-scale and multi-peak sample modeling. An adaptive boundary vector extraction algorithm is used to extract the boundary vectors which include all support vectors from the training samples, so that reduces the sample scale. The global optimal regression model is obtained by solving the multiscale v-SVM quadratic programming problems, and the complex distribution sample can be approximated from multiple scales by the model. Simulation results show that the v-support vector machine based on boundary vector extraction has better regression results than the general v-SVM.

Keywords: large sample modeling; boundary vector extraction; multiscale study; v-support vector machine

0 引 言

长期以来, 系统建模一直被视为解决实际问题的有效途径. 但是, 如何快速地建立高精度模型始终是系统建模研究的核心问题[1]. 支持向量机 (SVM) 是由 Vapnik[2]提出的一种新型系统建模方法, 由于其具有良好的小样本学习和泛化能力, 已成为建模方法的重要研究方向, 并在许多领域得到了成功应用[3]. 但是, SVM 在处理某些问题时也表现出不足, 例如在实际中很多数据都呈现大规模、多峰的特点, 当利用这类数据进行训练建模时, 普通 SVM 往往不能够很好地完成建模任务, 极易出现建模速度慢、回归精度低的现象. 为了弥补普通 SVM 的这种局限性, 一些学者运用聚类 SVM 等方法[4-6]成功地提高了 SVM 的训练速度, 但采取这些方法进行建模时很可能导致训练数据“信息”的丢失, 很大程度上影响了模型的精度. 与此同时, 还有一些学者对支持向量机的核函数进行改造, 提出了超核函数等核方法[7-9], 虽然这些方法在一些应用领域能够取得满意的回归效果, 但是运用这类单尺度核对不平坦分布样本进行建模时, 依然不能够得到满意的回归精度. 可以看出, 上述方法都只能单独解决训练速度过慢或回归精度低的问题, 单纯利用这些方法并不能快速地建立准确的模型.

针对上述问题, 本文提出一种基于自适应边界向量提取的多尺度 v-支持向量机 (ABVE-Mv-SVM). 首先, 运用自适应边界向量提取方法, 从训练样本中提取出包含数据全部特征的边界样本集作为新的训练

收稿日期: 2014-01-08；**修回日期:** 2014-04-24.

基金项目: 国家自然科学基金项目(61203021)；辽宁省科技攻关项目(2011216011).

作者简介: 苏成利(1977−), 男, 教授, 从事预测控制和工业过程先进控制及优化等研究；郑博元(1989−), 男, 硕士生, 从事机器学习、数据挖掘的研究.

样本, 在保证提取原始样本所有特征的同时限制边界样本集的规模, 以提高训练速度; 然后, 根据多尺度学习方法和SVM原理, 求解出多尺度v-SVM回归模型, 从多个尺度上消除回归残差, 得到更精确的模型.

1 问题描述

由支持向量机原理可知, 对于给定的训练集$(x_i, y_i)_{i=1,2,\cdots,n}$, $x_i \in R^l$为输入向量, $y_i \in R$为输出目标, 非线性SVM回归的目标就是寻找一个能够逼近y_i的拟合函数

$$f(x) = \sum_{i=1}^{n}(a_i^* - a_i)K(x, x_i) + b. \tag{1}$$

其中: a_i^*和a_i为拉格朗日常数, $K(x, x_i)$为核函数, b为偏差. 可以看出, 之所以式(1)具有非线性的表述能力, 是因为在RKHS上利用了核函数对数据进行逼近, 从而核函数的形式和结构便成为了影响拟合函数回归精度的最大因素. 实际应用中, 对于一般平稳的分布数据, 采用单个RBF核就可以获得良好的回归精度, 即

$$K(x, x_i) = \exp\Big(\frac{\|x - x_i\|}{2\sigma}\Big). \tag{2}$$

其中σ为尺度因子, 也称作核宽度. 若使用这种方法对多峰分布的数据进行逼近, 则很难使得此核函数能够以任意的精度逼近数据. 为了有效地解决这个问题, 文献[10]提出了一种可以从多个尺度上逐步消除回归残差的算法, 其具体算法如下:

在尺度1上对给定样本集$(x_i, y_i)_{i=1,2,\cdots,n}$进行回归, 模型为

$$f_1(x) = \sum_{i=1}^{m_1} w_{1i}K_1(x, x_i) + b_1. \tag{3}$$

其中: m_1为支持向量个数, $K_1(x, x_i)$为核函数, w_{1i}为相关核项系数, b_1为残差. 尺度2上的回归目标为给定样本输出y_i与尺度1输出$f_1(x_i)$二者之间的残差, 即$\{y_i - f_1(x_i)\}_{i=1,2,\cdots,n}$. 尺度2上的模型为

$$f_2(x) = \sum_{i=1}^{m_2} w_{2i}K_2(x, x_i) + b_2. \tag{4}$$

依据上述逐步消除残差的思想, 假设直到尺度m才能达到设定的逼近精度, 则最终模型为

$$f(x) = \sum_{k=1}^{m}\sum_{i=1}^{m_k} w_{ki}K_k(x, x_i) + b, \tag{5}$$

其中$b = \sum_{k=1}^{m} b_k$.

值得注意的是, 这种方法在获取高回归精度的同时也会造成训练速度慢的问题, 并且随着数据规模的增大, 此问题愈加凸显. 因为SVM的回归结果只会受到支持向量的影响, 所以若能找到一种能够在保存样本全部支持向量的同时又能缩减样本规模的方法, 则无疑会弥补多尺度学习造成的训练速度慢的问题.

2 自适应非线性边界向量提取算法

为了找到一种有效的数据规模约减方法, 首先引入以下定义.

定义 1 第i类样本$i \in \{1, 2\}$的平均特征为该类样本的样本中心m_i, 即

$$m_i = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{6}$$

定义 2 [11] 将m_1到m_2的方向$\overrightarrow{m_1m_2}$定义为第1类模式的特征方向(特征空间), m_2到m_1的方向$\overrightarrow{m_2m_1}$定义为第2类模式的特征方向(特征空间).

引理 1 [11] 已知2个向量$x = (x_1, x_2, \cdots, x_n), y = (y_1, y_2, \cdots, y_n)$, 经过非线性映射$\varphi(\cdot)$作用映射到特征空间中, 则这2个向量在特征空间中的Euclidean距离为

$$d^H(x, y) = \sqrt{K(x, x) - 2K(x, y) + K(y, y)}, \tag{7}$$

其中$K(\cdot, \cdot)$为核函数.

引理 2 [12] 已知3个向量$x = (x_1, x_2, \cdots, x_n), y = (y_1, y_2, \cdots, y_n)$及$z = (z_1, z_2, \cdots, z_n)$, 经过非线性映射$\varphi(\cdot)$的作用映射到特征空间中, 则在特征空间中, 向量$\varphi(x)\varphi(z)$在向量$\varphi(x)\varphi(y)$上的投影$\varphi(x)\varphi(z^o)$的Euclidean距离为

$$\|\varphi(x)\varphi(z^o)\|_2 = \frac{K(z, y) - K(z, x) - K(x, y) + K(x, x)}{\sqrt{K(x, x) - 2K(x, y) + K(y, y)}}. \tag{8}$$

根据上述理论, 有如下结论.

已知样本中心距离$d = \|\varphi(m_1)\varphi(m_2)\|_2$, 分别计算特征距离$\|\overrightarrow{\phi(m_1)\phi(x_i^o)}\|_2$和$\|\overrightarrow{\phi(m_2)\phi(y_j^o)}\|_2$, 令

$$r_1 = \max\|\overrightarrow{\phi(m_1)\phi(x_i^o)}\|_2, \tag{9}$$

$$r_2 = \max\|\overrightarrow{\phi(m_2)\phi(y_j^o)}\|_2. \tag{10}$$

引入非负修正因子$\delta \geqslant 0$.

1) 当$r_1 + r_2 < d$时, 若特征距离满足

$$r_1 - \delta \leqslant \|\overrightarrow{\phi(m_1)\phi(x_i^o)}\|_2 \leqslant r_1, \tag{11}$$

$$r_2 - \delta \leqslant \|\overrightarrow{\phi(m_2)\phi(y_j^o)}\|_2 \leqslant r_2, \tag{12}$$

则定义该模式为边界向量. 此时边界向量提取如图1所示, m_1表示第1类样本的中心, m_2表示第2类样本的中心. 边界向量预选取区域是直线D_1与直线r_1之间的部分和直线D_2与直线r_2之间的部分.

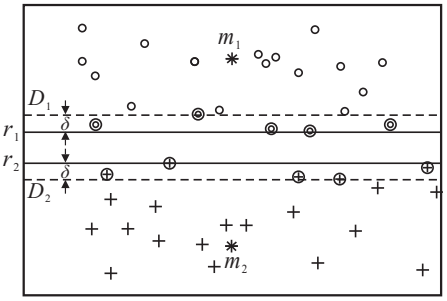

图 1 $r_1 + r_2 < d$时边界向量提取

2) 当 $r_1 + r_2 \geqslant d$ 时, 若特征距离满足

$$d - r_2 - \delta \leqslant \|\overrightarrow{\phi(m_1)\phi(x_i^o)}\|_2 \leqslant r_1 + \delta, \tag{13}$$

$$d - r_1 - \delta \leqslant \|\overrightarrow{\phi(m_2)\phi(y_j^o)}\|_2 \leqslant r_2 + \delta, \tag{14}$$

则定义该模式为边界向量. 此时边界向量提取如图 2 所示, m_1 为第 1 类样本中心, m_2 为第 2 类样本中心. 边界向量预选取区域是直线 $D_{1-\mathrm{up}}$ 与直线 $D_{1-\mathrm{down}}$ 之间的部分和直线 $D_{2-\mathrm{up}}$ 与 $D_{2-\mathrm{down}}$ 之间的部分.

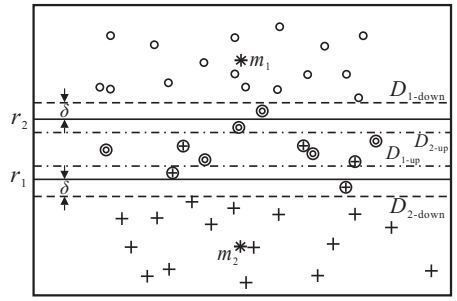

图 2　$r_1 + r_2 \geqslant d$ 时边界向量提取

由图 1 和图 2 可以看出, 边界向量集合的规模远远小于提取前的数据规模. 因为 SVM 算法的复杂度依赖于样本数据的数目, 样本数据数目越小, 求解相应的二次规划问题越简单, 所以利用边界向量代替原训练集对 SVM 进行训练可以提高训练速度. 但是, 上述理论并不完善, 修正参数 δ 的取值需要依据数据的分布情况来确定, 这意味着必须对数据的分布有深入的了解, 否则将很难获取满意的边界样本集合. 针对这一问题, 本文在上述理论的基础上提出了自适应提取边界向量算法.

为了使本算法能够有效地从原始样本中提取出所有支持向量, 本文定义自适应提取边界向量算法迭代停止条件如下:

$$\left|\frac{\mathrm{mse}(t-1) - \mathrm{mse}(t)}{\mathrm{mse}(t-1)}\right| \leqslant \eta. \tag{15}$$

其中: mse 代表经过边界向量训练的 SVM 产生的均方差, t 代表迭代次数. 一旦本次的 mse 与前一次 mse 的相对误差小于 η, 则本算法的迭代过程结束. 由此可知, η 的取值直接影响了算法的提取效果. 若 η 取值过小, 则易导致边界向量提取过度, 这意味着提取的边界向量集中包含了过多的“无用信息”; 如果 η 取值过大, 则易导致边界向量提取不足, 此时提取的边界向量集中并没有包含全部的支持向量, 导致数据的“信息缺失”. 考虑到以上因素, η 的取值范围一般为 $(0, 0.1)$.

自适应提取边界向量算法的具体步骤如下.

Step 1: 首先利用式 (6) 确定类中心, 再根据式 (8) ~ (10) 分别推导出特征空间中的距离 d 和特征空间中类最大特征距离 r_1、r_2.

Step 2: 设置自由参数 η 的初值、步长、范围, 并根据边界向量选取条件 (11)、(12) 或 (13)、(14) 将边界样本从原始样本中提取出来.

Step 3: 利用 Step 2 中提取的边界样本训练 SVM, 并将训练产生的 mse 代入式 (15). 当进行第 t 次迭代时, 若式 (15) 不成立, 则返回 Step 2, 按步长改变参数 δ, 重新提取边界样本; 如果式 (15) 成立, 则退出循环, 并将进行第 t 次迭代时的边界样本作为边界向量.

Step 4: 退出循环后, 若获得的边界向量数量较多, 没有达到期望的边界样本规模, 则适当增大 η 的值; 如果利用边界向量训练 v-SVM 导致预测精度下降, 则适当减小 η 的取值, 然后返回 Step 2 重新计算.

由上述步骤可以看出: 通过引入参数 η, 使得本算法可以在不了解数据分布的情况下提取出数据的全部特征, 并对边界向量集的规模进行了限制; 消除了 δ 的选取不当对提取结果造成的不良影响, 提升了算法性能.

3　多尺度 v-SVM

现有的多尺度 SVM 算法大都将不同核宽度的 SVM 进行线性组合来提高模型的精度[13], 但这类方法只能称为多尺度核 SVM, 不是真正的多尺度学习算法. 为了使 SVM 本质上具有多尺度特性的函数逼近能力, 文献 [14] 提出了多尺度 ε-SVM 算法, 但此方法在计算之前需要事先根据样本的分布来确定不敏感损失参数 ε. 当数据分布平坦时选择合适的 ε 较为简单, 但当样本分布复杂时, 选择合适的 ε 并不容易, 并且会对结果产生很大的影响. 由 v-SVM 原理[15]可知, v-SVM 可以通过控制 v 的取值来控制支持向量的个数和错误样本的个数, 其中 $v \in (0, 1)$. 相对于 ε-SVM, v-SVM 使用参数 v 替代 ε-SVM 中的参数 ε, 使参数的选取有了理论依据. 因此, 本文将多尺度方法与 v-SVM 相结合, 提出了多尺度 v-SVM 学习方法. 算法如下.

以双尺度为例. 对于给定样本 $(x_i, y_i)_{i=1,2,\cdots,n}$, 首先用大尺度核拟合平滑区域的样本, 大尺度上的回归模型为

$$f_1(x) = w_1 K_1(\cdot) + b_1, \tag{16}$$

其中 $K_1(\cdot)$ 为核函数; 然后用小尺度核拟合变化剧烈区域的样本,目标为 $\{(x_i, y_i) - f_1(x_i)\}_{i=1,2,\cdots,n}$, 则小尺度上的回归模型为

$$f_2(x) = w_2 K_2(\cdot) + b_2, \tag{17}$$

最终的双尺度回归模型为

$$f(x) = w_1 K_1(\cdot) + b_1 + w_2 K_2(\cdot) + b_1 + b_2. \tag{18}$$

根据多尺度学习思想和 SVM 原理, 若欲求得式 (18) 的解, 则可通过计算下面的优化问题求出:

$$\min \frac{1}{2}\|w_1\|^2 + \frac{1}{2}\|w_2\|^2 + \frac{C_1}{n}\sum_{i=1}^{n}(\xi_{1i} + \xi_{1i}^*) +$$

$$\frac{C_2}{n}\sum_{i=1}^{n}(\xi_{2i}+\xi_{2i}^*)+C_1v_1\varepsilon_1+C_2v_2\varepsilon_2;$$

$$\text{s.t. } w_1\phi_1(x_i)+b_1-y_i\leqslant\varepsilon_1+\xi_{1i},$$

$$y_i-w_1\phi_1(x_i)-b_1\leqslant\varepsilon_1+\xi_{1i}^*,$$

$$w_2\phi_2(x_i)+b_2-(y_i-w_1\phi_1(x_i)-b_1)\leqslant$$

$$\varepsilon_2+\xi_{2i},$$

$$(y_i-w_1\phi_1(x_i)-b_1)-w_2\phi_2(x_i)-b_2\leqslant$$

$$\varepsilon_2+\xi_{2i}^*. \tag{19}$$

其中: C_1 和 C_2 为惩罚因子, ξ_{1i}、ξ_{1i}^*、ξ_{2i}、ξ_{2i}^* 为松弛系数, v_1、v_2 和 ε_1、ε_2 分别为尺度1和尺度2上的参数. 构造拉格朗日函数为

$$\min\ L=$$

$$\frac{1}{2}\|w_1\|^2+\frac{1}{2}\|w_2\|^2+\frac{C_1}{n}\sum_{i=1}^{n}(\xi_{1i}+\xi_{1i}^*)+$$

$$\frac{C_2}{n}\sum_{i=1}^{n}(\xi_{2i}+\xi_{2i}^*)+C_1v_1\varepsilon_1+C_2v_2\varepsilon_2-$$

$$\beta_1\varepsilon_1-\beta_2\varepsilon_2-\sum_{i=1}^{n}\alpha_{1i}(\varepsilon_1+\xi_{1i}+y_i-$$

$$w_1\phi_1(x_i)-b_1)-\sum_{i=1}^{n}\alpha_{1i}^*(\varepsilon_1+\xi_{1i}^*-y_i+$$

$$w_1\phi_1(x_i)+b_1)-\sum_{i=1}^{n}\alpha_{2i}(\varepsilon_2+\xi_{2i}-$$

$$w_2\phi_2(x_i)-b_2+y_i-w_1\phi_1(x_i)-b_1)-$$

$$\sum_{i=1}^{n}\alpha_{2i}^*(\varepsilon_2+\xi_{2i}^*+w_2\phi_2(x_i)+$$

$$b_2-y_i+w_1\phi_1(x_i)+b_1)-$$

$$\sum_{i=1}^{n}(\eta_{1i}\xi_{1i}+\eta_{1i}^*\xi_{1i}^*+\eta_{2i}\xi_{2i}+\eta_{2i}^*\xi_{2i}^*), \tag{20}$$

其中 $\alpha_{1i}^{(*)}$、$\alpha_{2i}^{(*)}$ 为拉格朗日乘子. 分别对 w_1、w_2、b_1、b_2、ε_1、ε_2、$\xi_{1i}^{(*)}$、$\xi_{2i}^{(*)}$ 求偏导, 并令其结果都等于0, 可得式(19)的对偶问题

$$\max\ W=\sum_{i=1}^{n}y_i(\alpha_{1i}^*-\alpha_{1i}+\alpha_{2i}^*-\alpha_{2i})-$$

$$\frac{1}{2}\sum_{i=1,j=1}^{n}[(\alpha_{1i}^*-\alpha_{1i}+\alpha_{2i}^*-\alpha_{2i})(\alpha_{1j}^*-$$

$$\alpha_{1j}+\alpha_{2j}^*-\alpha_{2j})K_1(x_i,x_j)+$$

$$(\alpha_{2i}^*-\alpha_{2i})(\alpha_{2j}^*-\alpha_{2j})K_2(x_i,x_j)];$$

$$\text{s.t. }\sum_{i=1}^{n}(a_{1i}-a_{1i}^*+a_{2i}-a_{2i}^*)=0,$$

$$\sum_{i=1}^{n}(a_{2i}-a_{2i}^*)=0,$$

$$0\leqslant\alpha_{1i}^{(*)}\leqslant\frac{C_1}{n},\ 0\leqslant\alpha_{2i}^{(*)}\leqslant\frac{C_2}{n},$$

$$\sum_{i=1}^{n}(\alpha_{1i}+\alpha_{1i}^*)\leqslant C_1v_1,$$

$$\sum_{i=1}^{n}(\alpha_{2i}+\alpha_{2i}^*)\leqslant C_2v_2. \tag{21}$$

由式(21)可见, 式中并没有不敏感损失参数 ε, 这是由于在式(20)到(21)的计算过程中包含 ε 的项被约去了, 这说明 ε 的取值不会对最终结果产生影响. 在求得拉格朗日乘子 $\alpha_{1i}^{(*)}$、$\alpha_{2i}^{(*)}$ 的最优解后, 可得到决策函数

$$f(x)=\sum_{i=1}^{n}(\alpha_{1i}^*-\alpha_{1i}+\alpha_{2i}^*-\alpha_{2i})K_1(x,x_i)+$$

$$\sum_{i=1}^{n}(\alpha_{2i}^*-\alpha_{2i})K_2(x,x_i)+b, \tag{22}$$

其中 $b=\dfrac{1}{n}\sum_{i=1}^{n}(y_i-f(x_i))$.

由上述推导可知, 在本算法获取双尺度 v-SVM 最优回归模型的过程中, 用参数 v 代替了参数 ε, 降低了参数选择的难度, 提高了算法对不平坦分布数据的适应性. 对于更多尺度 v-SVM 算法, 可利用上述方法进行推广, 在此不再赘述.

4 仿真研究

本文的实验数据采用上海证券交易所上证指数数据 (Data source: Great wisdom securities piece). 因为该数据具有样本数量大、分布复杂的特点, 所以很适合检验本文提出的 ABVE-M v-SVM 方法的有效性, 并且该数据能比较准确地反映国内股市行情动态, 具有很高的建模价值. 该数据包含6个特征, 共4 579个样本. 为验证本文所提出方法的有效性, 将自适应边界向量提取 v-SVM (ABVE-v-SVM)、多尺度 v-SVM (Mv-SVM) 和 ABVE-M v-SVM 与 v-SVM 方法进行对比研究. 从上海证券交易所上证指数数据集中生成一个1 100个数据的样本集, 其输入为开盘指数、指数最高值、指数最低值、收盘指数、当日交易量, 输出为当日交易额. 其中1 000个样本作为训练数据集, 另外100个样本作为测试数据集. 采用 $\text{MSE}=\dfrac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y}_i)^2$ 作为预测精度指标, 时间 t (s) 作为训练速度指标. 其中: n 为样本总数; y_i 为测试样本输出; $\overline{y_i}$ 为预测样本预测输出, $i=1,2,\cdots,n$. 仿真研究中采用实验法对参数进行寻优, 具体过程如表1～表3所示.

表 1 尺度1核参数 σ_1 选择过程

σ_1	σ_2	C	MSE
7	1.5	50	$7.585\,6\times10^{15}$
6	1.5	50	$3.684\,4\times10^{15}$
5	1.5	50	$2.850\,2\times10^{15}$
4	1.5	50	$4.856\,0\times10^{15}$
3	1.5	50	$5.914\,8\times10^{15}$

表 2　尺度 2 核参数 σ_2 选择过程

σ_1	σ_2	C	MSE
5	1.2	50	$2.705\,3 \times 10^{15}$
5	1.1	50	$2.235\,9 \times 10^{15}$
5	1.0	50	$1.603\,8 \times 10^{15}$
5	0.9	50	$2.363\,4 \times 10^{15}$
5	0.8	50	$2.540\,4 \times 10^{15}$

表 3　惩罚参数 C 选择过程

σ_1	σ_2	C	MSE
5	1.0	45	$2.405\,3 \times 10^{15}$
5	1.0	40	$1.022\,5 \times 10^{15}$
5	1.0	35	$3.260\,3 \times 10^{14}$
5	1.0	30	$5.458\,2 \times 10^{14}$
5	1.0	25	$7.297\,5 \times 10^{14}$

根据均方差最小原则, 通过比较表 1 ~ 表 3 中的回归结果, 最终确定 $C = 35$, $\sigma_1 = 5$, $\sigma_2 = 1.0$. 同理, 依据上述的参数选择方法可确定 $v = 0.5$, $\eta = 0.05$. 仿真结果如图 3 ~ 图 6 所示.

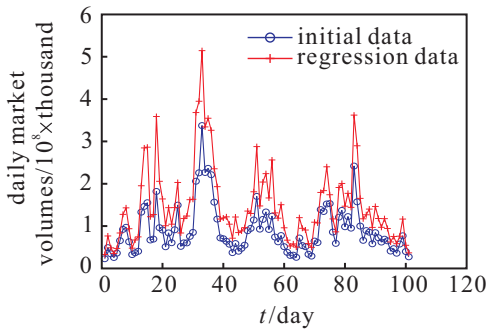

图 3　v-SVM 测试集回归结果

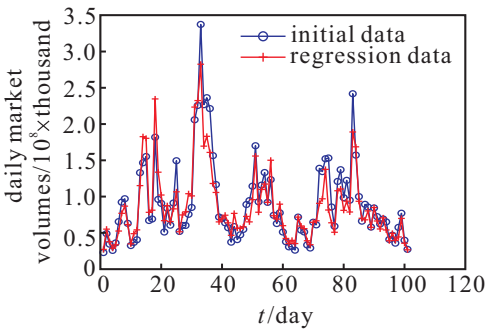

图 4　ABVE-v-SVM 测试集回归结果

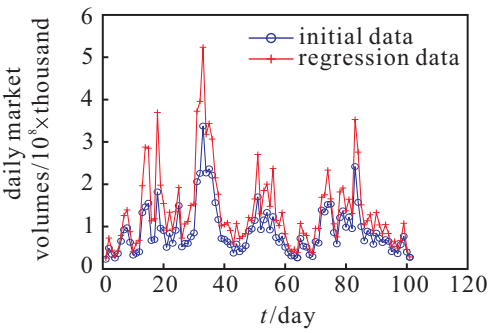

图 5　M v-SVM 测试集回归结果

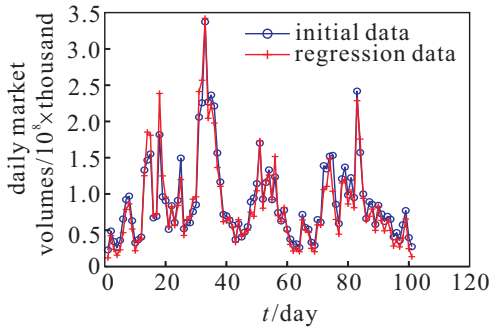

图 6　ABVE-M v-SVM 测试集回归结果

图 3 为 v-SVM 方法的测试集回归结果, 图 4 为 ABVE-v-SVM 方法的测试集回归结果. 通过对比图 3 和图 4 可以看到, 与 v-SVM 的回归曲线相比, ABVE-v-SVM 回归曲线并没有因为训练集规模的缩小而受到很大影响. 表 4 为 v-SVM 与 ABVE-v-SVM 训练速度比较, 可以看出 ABVE-v-SVM 的支持向量数量与训练时间均小于 v-SVM. 上述结果说明, ABVE 算法在不影响回归精度的前提下, 能够从大规模样本中提取含有全部支持向量的边界向量集, 能够有效削减训练样本的规模, 提升训练速度.

表 4　v-SVM 与 ABVE-v-SVM 训练速度比较

训练算法	训练样本	边界向量数量	支持向量数量	时间/s
v-SVM	1 000	无	261	21
ABVE-v-SVM	1 000	113	25	2

下面将 M v-SVM 与 v-SVM 进行对比, 图 3 和图 5 分别为 v-SVM 测试集回归结果和 M v-SVM 预测集回归结果. 通过对比可以看出, M v-SVM 的预测精度明显高于 v-SVM 的预测精度. 这说明本文提出的 M v-SVM 能够适应不均匀的样本学习问题, 并有效地提高了回归精度.

图 6 表示的是 ABVE-M v-SVM 测试集回归结果, 与图 3 ~ 图 5 相比, 图 6 的回归曲线显然具有更好的拟合效果.

由表 5 可以看出: 与 v-SVM 相比, 在具有相近回归精度的情况下, ABVE-v-SVM 的训练时间更短, 这说明 ABVE 能够在不影响回归精度的前提下对训练样本的规模进行缩减; M v-SVM 与 v-SVM 相比较, M v-SVM 获得了更小的 MSE, 这体现 M v-SVM 能够从多个尺度上进行回归, 有效地提高了回归的精度. 通过将 ABVE-M v-SVM 与 v-SVM 的比较可以发现, ABVE-M v-SVM 不仅具有更短的训练时间, 而且获得了更高的回归精度. 由上述对比可知, ABVE-M v-SVM 兼顾了多尺度学习和 ABVE 的优点, 并克服了它们相应的缺点, 运用此方法能有效地解决普通 v-SVM 用于大规模、多峰数据建模时的不足.

表 5 回归结果比较

训练算法	训练样本	训练时间/s	核参数	MSE
v-SVM	1 000	21	$\sigma = 5$	4.708×10^{15}
ABVE-v-SVM	1 000	2	$\sigma = 5$	4.713×10^{15}
Mv-SVM	1 000	35	$\sigma_1 = 5, \sigma_2 = 1.0$	3.341×10^{14}
ABVE-M v-SVM	1 000	3	$\sigma_1 = 5, \sigma_2 = 1.0$	3.260×10^{14}

5 结 论

本文通过分析普通v-SVM对大规模、多峰数据进行回归时存在的问题和原因, 并结合边界向量提取算法和多尺度学习方法, 提出了基于自适应边界向量提取的多尺度v-SVM, 并将该方法与普通v-SVM进行了比较. 仿真研究结果表明, 本文算法有以下两个优点: 一是可以有效地缩减样本的规模, 使得建模速度有很大的提升; 二是能够在多个尺度上对复杂分布样本进行逼近, 提高模型的回归精度. 上述优点说明ABVE-Mv-SVM建模方法兼备了快速性和精确性, 具有广阔的应用前景.

参考文献(References)

[1] 钱学森, 于景元, 戴汝为. 开放的复杂巨系统及其方法论[J]. 城市发展研究, 2005, 12(5): 1-8.
(Qian X S, Yu J Y, Dai R W. The study of open complex giant system and its methodology[J]. Urban Studies, 2005, 12(5): 1-8.)

[2] Vapnik V N. The nature of statistical learning theory[M]. New York: Spring-Verlag, 1995: 123-167.

[3] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42.
(Zhang X G. Introduction to statistical learning theory and support vector machines[J]. Acta Automatica Sinica, 2000, 26(1): 32-42.)

[4] Shi Jinhong, Su Mingyang, Chen Yuansi. A novel intrusion detection system based on hierarchical clustering and support vector machines[J]. Expert Systems with Application, 2011, 38(1): 175-181.

[5] Niu Dongxiao, Wei Yanan. An improved short-term power load combined forecasting with ARMA-GRACH-ANN-SVM based on FHNN similar-day clustering[J]. J of Software, 2013, 8(3): 341-346.

[6] 丁志勇, 杨苹, 杨曦, 等. 基于连续时间段聚类的支持向量机风电功率预测方法[J]. 电力系统自动化, 2012, 36(14): 131-135.
(Ding Z Y, Yang P, Yang X, et al. Wind power prediction method based on sequential time clustering support vector machine[J]. Automation of Electric Power Systems, 2012, 36(14): 131-135.)

[7] 陈强, 任雪梅. 基于多核最小二乘支持向量机的永磁同步电机混沌建模及其实时在线预测[J]. 物理学报, 2010, 59(4): 2311-2318.
(Chen Q, Ren X M. Chaos modeling and real-time online prediction of permanent magnet synchronous motor based on multiple kernel least squares support vector machine[J]. Acta Physica Sinica, 2010, 59(4): 2311-2318.)

[8] 周金柱. 集成先验知识的多核线性规划支持向量回归[J]. 自动化学报, 2011, 37(3): 360-370.
(Zhou J Z. Multiple kernel linear programming support vector regression incorporating prior knowledge[J]. Acta Automatica Sinica, 2011, 37(3): 360-370.)

[9] Cai Yanning, Wang Hongqiao. A multiple-kernel LSSVR method for separable nonlinear system identifcation[J]. J of Control Theory and Applications, 2013, 11(4): 651-655.

[10] 汪洪桥, 蔡艳宁, 孙富春, 等. 多尺度核方法的自适应序列学习及应用[J]. 模式识别与人工智能, 2011, 24(1): 72-81.
(Wang H Q, Cai Y N, Sun F C, et al. Adaptive sequence learning and applications for multi-scale kernel method[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(1): 72-81.)

[11] 焦李成, 张莉, 周伟达. 支撑向量预选取的中心距离比值法[J]. 电子学报, 2001, 29(3): 383-386.
(Jiao L C, Zhang L, Zhou W D. Center distance ratio method support vector pre-selection[J]. Acta Electronica Sinica, 2001, 29(3): 383-386.)

[12] 李青, 焦李成. 基于向量投影的支撑向量预选取[J]. 计算机学报, 2005, 28(2): 145-151.
(Li Q, Jiao L C. Pre-extracting support vector for support vector machine based on vector projection[J]. Chinese J of Computer, 2005, 28(2): 145-151.)

[13] Kim M. Accelerated max-margin multiple kernel learning[J]. Applied Intelligence, 2013, 38(1): 45-57.

[14] 任世锦, 吴铁军. 基于径向基小波核的多尺度小波支持向量机[J]. 电路与系统学报, 2008, 13(4): 70-75.
(Ren S J, Wu T J. Multi-scale wavelet SVM based on radial wavelet kernel[J]. J of Circuits and Systems, 2008, 13(4): 70-75.)

[15] 吴奇, 严洪森. 基于混沌v-支持矢量机的产品销售预测模型[J]. 机械工程学报, 2010, 46(7): 128-135.
(Wu Q, Yan H S. Forecasting model of product sales based on the chaotic v-support vector machine[J]. J of Mechanical Engineering, 2010, 46(7): 128-135.)

(责任编辑: 曹洪武)