文章编号: 1001-0920(2011)02-0316-04

基于自适应增强算法的支持向量机组合模型

杨慧中, 邓玉俊

(江南大学 通信与控制工程学院, 江苏 无锡 214122)

摘 要: 为了提高软测量模型的泛化能力, 提出一种基于 AdaBoosting 算法的组合支持向量机 (SVM) 模型. 该方法在贝叶斯分析的基础上, 利用样本概率初始化惩罚系数, 依据回归过程中的损失函数更新惩罚系数权重, 使得 SVM 训练模型有强、弱之分, 突出一些重要样本的作用, 以提高模型的估计精度和泛化能力. 仿真结果表明, 依据该方法建立的组合模型明显改善了软测量模型的估计能力和泛化能力.

关键词: 支持向量机; 自适应增强算法; 组合模型

中图分类号: TP18; TP274 **文献标识码:** A

Compositional model of SVM based on AdaBoosting algorithm

YANG Hui-zhong, *DENG Yu-jun*

(School of Communication and Control Engineering, Jiangnan University, Wuxi 214122, China. Correspondent: YANG Hui-zhong, E-mail: yhz_jn@163.com)

Abstract: In order to improve the generalization ability of a soft-sensor model, a compositional model of SVM based on AdaBoosting algorithm is proposed. On the basis of Bayesian analysis, the penalty coefficient is initialized by using the Bayesian probability of the samples, and then the penalty weight is updated by the loss function in the regression process so that the SVM training model can highlight some important samples to improve its estimation accuracy and generalization ability. Simulation result shows that this approach can greatly improve the estimation capacity and generalization ability of the model.

Key words: support vector machine; AdaBoosting algorithm; compositional model

1 引 言

软测量技术为复杂工业生产过程中难以用传统仪表直接测量参数提供了一种有效的解决方法, 但复杂工业过程的多变量、非线性、强耦合、时变时滞以及不确定等特性, 给软测量模型的建模与辨识带来了巨大挑战. 基于多模型策略的建模方法, 采用多个子模型来逼近过程对象特性以实现精确估计是当前解决复杂过程对象建模的有效方案. 但多模型的建立依然建立在各个子模型的基础上, 每一个子模型的估计精度得不到很好的提高, 导致最终组合模型的估计精度不能得到较好的改善.

为了进一步提高单个支持向量机 (SVM) 的泛化精度, 本文在学习 AdaBoosting 基本算法[1-2]的基础上, 将其从分类算法中应用到回归模型中, 并提出一种新的损失函数的定义方法: 利用最大相对误差和均方差定义损失函数. 对最大相对损失函数进行回归来控制模型对突变数据的跟踪, 均方差用来控制整个模型的估计趋势, 使得模型不会因为跟踪某一突变数据而使整个估计趋势发生变化. 该算法的主要目的是提高单个 SVM 的回归性能, 针对不同样本设定不同的惩罚系数, 将弱 SVM 经过算法提升后变成强 SVM, 从而提高模型的估计精度. 工业过程仿真表明, 基于贝叶斯分类器和 Boosting 算法的组合模型 (SVM-BC-B) 的估计精度要高于基于贝叶斯分类器的组合 SVM 模型 (SVM-BC).

2 基本原理

2.1 朴素贝叶斯分类器

朴素贝叶斯分类器 (NBC)[3-4]是贝叶斯分类模型中一种简单有效且在实际应用中比较成功的分类器. 贝叶斯分类器的目的是在给定描述对象的特征值

收稿日期: 2009-10-27; **修回日期:** 2010-01-14.

基金项目: 国家自然科学基金项目(60674092).

作者简介: 杨慧中(1955–), 女, 教授, 博士生导师, 从事过程建模、优化控制等研究; 邓玉俊(1984–), 男, 硕士生, 从事数据挖掘技术的研究.

$\{x_1, x_2, \cdots, x_n\}$ 下, 得到最有可能的目标值 C_{NBC}, 即

$$C_{\mathrm{NBC}} = \arg\max P(c_j/x_1, x_2, \cdots, x_n), \quad (1)$$

其中 $j = 1, 2, \cdots, m$ 为类别标号. 根据贝叶斯公式, 式 (1) 可以改写为

$$C_{\mathrm{NBC}} = \arg\max \frac{P(x_1, x_2, \cdots, x_n/c_j)}{P(x_1, x_2, \cdots, x_n)}. \quad (2)$$

其中: $P(x_1, x_2, \cdots, x_n/c_j)$ 为样本 $(x_1, x_2, \cdots, x_n)$ 属于类别 c_j 的条件概率, $P(x_1, x_2, \cdots, x_n)$ 为样本 $(x_1, x_2, \cdots, x_n)$ 的联合概率. 估计 $P(c_j)$ 较为容易, 只需计算出每个样本类别在训练数据中的频率即可. 然而, 估计 $P(x_1, x_2, \cdots, x_n/c_j)$ 却较难实现, 除非有一个比较大的样本集.

朴素贝叶斯分类器基于一个简单的假定: 在给定目标值时属性相互之间条件独立. 该假定说明在给定实例目标值的情况下, 观察到 $\{x_1, x_2, \cdots, x_n\}$ 的联合概率等于单个属性的概率乘积, 即

$$C_{\mathrm{NBC}} = P(c_j/x_1, x_2, \cdots, x_n) = \frac{P(x_1, x_2, \cdots, x_n|c_j) \times P(c_j)}{P(x_1, x_2, \cdots, x_n)} \propto P(c_j)\prod_{i=1}^{n} P(x_i|c_j). \quad (3)$$

将式 (3) 代入 (2), 可得到朴素贝叶斯分类器所使用的方法. 朴素贝叶斯分类器输出的目标值为

$$C_{\mathrm{NBC}} = \arg\max P(c_j)\prod_{i=1}^{n} P(x_i|c_j). \quad (4)$$

2.2 Boosting 算法

Boosting 算法由 Schapire 提出, 最初用在分类中, 其原理是在一个线性决策规则中组合几种弱学习算法, 这样组合的线性决策规则比任何一个弱学习算法的性能均好得多. 弱学习算法是指一个学习效果比随机分类略好的学习算法. Boosting 算法的基本思想是对那些容易分错类的训练样本加强学习, 具体步骤如下[5]:

Step 1: 将每个训练样本赋予相同的权重, 训练第 1 个基本分类器, 并用它来对训练集进行测试;

Step 2: 对于那些分类错误的测试样本提高其权重, 用调整后的带权训练集训练第 2 个基本分类器;

Step 3: 重复该过程, 直到得到一个足够好的学习器.

Boosting 的基本目标是将一个弱学习器转化为一个任意的高精度学习器. 在众多 Boosting 算法中, AdaBoosting 是最具有实用价值的. 它根据弱学习机的反馈, 自适应地调整学习机的错误率: 如果前次循环生成的弱学习机频繁地在某一样本上发生分类错误, 则该样本将被赋予较大的权重; 在下一轮循环中, 弱算法会将注意力集中到权重较大且分类困难的样本上, 从而达到较高的泛化精度. 文献 [5] 将其应用推广到回归问题中, 通过将回归问题转化为多个分类问题来研究. 虽然思路较新颖, 但增大了计算量, 且泛化性能不是很好, 所以不是很实用.

3 改进的 AdaBoosting 支持向量回归算法

标准支持向量回归机对于训练数据中所有样本的惩罚程度均一样, 惩罚系数均为 C, 并没有突出对难学习样本的惩罚. 而在本文所提出的算法中, 其惩罚值是针对各个样本的, 样本 x_i 对应的惩罚值为 Cw_i. 该算法所对应的优化问题为

$$\begin{aligned} &\min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l} w_i(\xi_i + \xi_i^*),\ i = 1, 2, \cdots, l;\\ &\text{s.t. } y_i - \langle\omega, \phi(x_i)\rangle - b \leqslant \varepsilon + \xi_i,\\ &\qquad \langle\omega, \phi(x_i)\rangle + b - y_i \geqslant \varepsilon + \xi_i^*,\ \xi_i, \xi_i^* \geqslant 0. \end{aligned} \quad (5)$$

将其转化为二次规划后得到的对偶优化问题为

$$\begin{aligned} &\min \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle\phi(x_i), \phi(x_i)\rangle + \\ &\qquad \sum_{i=1}^{l}\alpha_i(\varepsilon_i - y_i) + \sum_{i=1}^{l}\alpha_i^*(\varepsilon_i^* + y_i);\\ &\text{s.t. } \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0,\ \alpha_i, \alpha_i^* \in [0, Cw_i]. \end{aligned} \quad (6)$$

其中 $w_i(\cdot)$ 为第 i 个样本对应的权值, 其初始值为样本概率 p_i. 这样可将权值的初始值与数据的特性相结合, 减少程序的循环次数. $w_i(\cdot)$ 的更新与前一次训练的结果有关, 即样本惩罚程度按其误差大小决定, 也就是要求误差大的样本对应的权重较大, 误差小的样本对应的权重较小.

为了得到 $w_i(\cdot)$ 的更新公式, 设训练样本 x_i 对应的绝对损失函数为

$$l_t(i) = \begin{cases} 0, & |f(x_i) - y_i| \leqslant \varepsilon;\\ |f(x_i) - y_i| - \varepsilon, & |f(x_i) - y_i| > \varepsilon. \end{cases} \quad (7)$$

其对应的相对损失函数为

$$L(i) = \begin{cases} 0, & |f(x) - y| \leqslant \varepsilon;\\ \dfrac{|f(x_i) - y_i| - \varepsilon}{y_i}, & |f(x) - y| > \varepsilon. \end{cases} \quad (8)$$

则整个模型的损失函数定义为其对应的相对损失函数, 即

$$\mathrm{Loss} = \frac{L_m}{\sum_{i=1}^{n} L(i)} + \sqrt{\frac{1}{n-1}\sum_{i=1}^{n-1} l(i)^2 w(i)^2}, \quad (9)$$

其中 $L_m = \max(L(i))$ 为样本集的最大相对损失函数. 由式 (9) 可知, 该损失函数包括两部分: 模型的相对损失函数和模型的均方损失函数. 定义这样的损失函数的目的是, 利用误差的反馈来调整模型的性能指标和

整体的估计趋势.

将支持向量回归机的权重系数 β_t 取为损失函数的函数, 令 $\beta_t = \mathrm{Loss}(t)^2$. 各训练样本所对应的权值为 $w_i(\cdot)$ 的更新, 根据回归机的权重系数及损失函数重新分配, 要求误差大的样本对应的权重较大, 即对其惩罚较大; 要求误差小的样本其对应的权重也较小, 即对其惩罚也较小, 则权值更新公式为

$$w_{t+1}(i) = w_t(i)\beta_t^{(1-L_t(i))}/D_t. \tag{10}$$

其中: $D_t = \sum_{i=1}^{l} w_{t+1}^i$, $w_0 = p$. 最终输出的学习回归机为

$$f(x) = \sum_t \left(\frac{1}{\beta_t}\right) f_t(x) \Big/ \sum_t \left(\frac{1}{\beta_t}\right).$$

由此可以看出, $1/\beta_t$ 实际上是对回归机 $f_t(x)$ 可信度的一种度量. β_t 较大, 其对应的 $f_t(x)$ 的平均误差较大, 在输入 $f(x)$ 中该 $f_t(x)$ 所占比例相对较小; 反之, β_t 较小, 其所对应的 $f_t(x)$ 在结果中所占的比例相对较大.

4 模型建立

利用上述方法建立基于改进的Boosting算法的支持向量机模型, 其建立步骤如下:

Step 1: 划分数据的训练集和测试集, 利用训练数据建立贝叶斯分类器, 得到每一个样本属于所属类别的概率;

Step 2: 利用样本概率初始化基于Boosting算法的SVM惩罚系数, 设定误差阈值;

Step 3: 训练类别模型, 计算损失函数, 进行惩罚系数权重的递归计算;

Step 4: 计算模型误差, 与误差阈值进行比较, 当小于阈值时停止循环; 否则重复Step 3, 直到小于误差阈值;

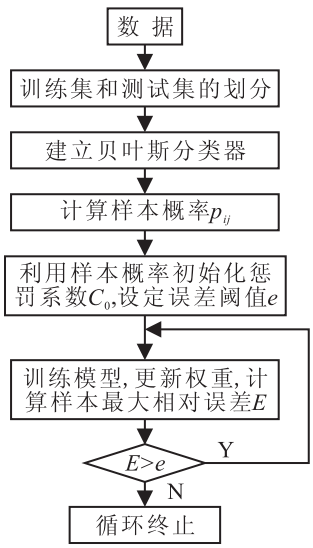

图 1 类别子模型的训练

Step 5: 重复Step 3和Step 4, 直到所有的类别模型均训练结束, 利用开关式的方式建立组合模型;

Step 6: 计算测试样本的类别概率, 判断其所属的类别, 利用测试数据对所属类别子模型进行检测.

图1描述了子模型训练的过程, 其中 p_{ij} 表示第 i 个样本属于第 j 类的概率. 图2给出了整个SVM-BC-B模型的建立结构.

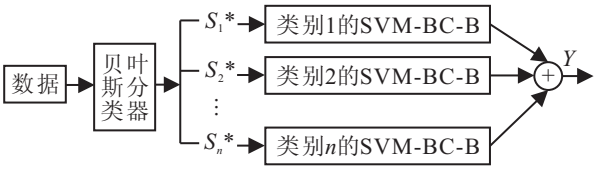

图 2 组合模型的建立

5 仿真研究

利用本文提出的基于Boosting算法的组合SVM模型对丙烯腈聚合反应生产过程中聚合物总数质量指标——平均分子量进行在线估计. 研究丙烯腈聚合反应工艺可知, 影响聚合釜聚合物总数的主要操作变量为亚硫酸浓度、引发剂与单体浓度之比、停留时间以及反应温度. 从现场采集的98组数据中选取66组作为训练样本, 32组作为测试样本. 训练样本经归一化等预处理后, 利用贝叶斯分类器对其进行分类, 得到3个类别. 计算每1个类别的概率, 以及每1个属性区域相对于类别的先验概率, 判断出样本的所属类别. 分别将每1个类别的数据输入到改进后的SVM模型中, 得到3个子模型, 对其利用开关式的方式进行组合, 得到基于Boosting算法的组合SVM模型. 图3为基于Boosting算法改进的SVM模型和利用标准SVM模型的测试比较图, 表1给出了两个模型估计的误差对比.

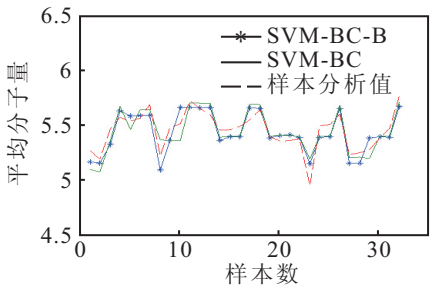

图 3 模型SVM-BC-B以及SVM-BC测试曲线

表 1 两种模型的误差表

model	maximum relative error/%	mean square deviation/%
SVM-BC	4.98	1.81
SVM-BC-B	3.92	1.61

从图3和表1可以看出, 利用Boosting算法改进子模型, 最终达到了改进组合模型估计精度的效果.

改进的Boosting算法不仅降低了模型的最大估计误差, 而且在一定程度上使模型的整体趋势更接近对象的输出.

6 结　论

虽然SVM-BC模型已经改善了SVM的估计精度, 但由于SVM自身存在的缺陷, 使模型的估计精度不能得到进一步的提高. SVM-BC-B模型从SVM自身的缺陷出发来进行改进, 从仿真的结果来看, 这种方法能够进一步提高SVM模型的估计精度, 具有一定的实用性.

参考文献(References)

[1] Freund Y, Schapire R. A decision-theoretic generalization of online learning and an application to boosting[J]. J of Computer and System Science, 1997, 55(1): 119-139.

[2] Freund Y, Schapire R. Experiments with a new boosting algorithm[C]. Machine Learning: Proc of the 13th Int Conf. Morgan Kaufmann, 1996: 148-156.

[3] Hall M. A decision tree-based attribute weighting filter for naive Bayes[J]. Knowledge-based Systems, 2007, 20(2): 120-126.

[4] Addin O, Sapuan S, Mahdi E, et al. A naive-bayesian classifier for damage detection in engineering materials[J]. Materials and Design, 2007, 28(8): 2379-2386.

[5] Ridgeway G, Madigan D, Richardson T. Boosting methodology for regression problems[C]. Proc of the 7th Int Workshop on Artificial Intelligence and Statistics. San Francisco: Morgan Kaufmann Publishers, 1999: 152-161.

(上接第315页)

[2] 邓聚龙. 灰理论基础[M]. 武汉: 华中科技大学出版社, 2002: 1-46.
(Deng J L. The foundation of grey system[M]. Wuhan: Huazhong University of Science and Technology Press, 2002: 1-46.)

[3] Deng J L. A novel GM(1,1) model for nonequigap series[J]. The J of Grey System, 1997, 9(2): 111-116.

[4] Xiao X P, Deng J L. A new modified GM(1,1) model: Grey optimization model[J]. The J of Systems Engineering and Electronics, 2001, 23(2): 1-5.

[5] 王义闹. GM(1,1)的直接建模方法及性质[J]. 系统工程理论与实践, 1988, 1(1): 27-31.
(Wang Y N. The direct method and its character of GM(1,1)[J]. Systems Engineering-Theory & Practice, 1988, 1(1): 27-31.)

[6] 张岐山. 提高灰色GM(1,1)模型精度的微粒群方法[J]. 中国管理科学, 2007, 15(5): 126-129.
(Zhang Q S. Improving the precision of GM(1,1) model by using particle swarm optimization[J]. Chinese J of Management Science, 2007, 15(5): 126-129.)

[7] 党耀国, 刘思峰, 刘斌. 以$x(1)(n)$为初始条件的GM(1,1)模型[J]. 中国管理科学, 2005, 13(1): 132-134.
(Dang Y G, Liu S F, Liu B. The GM models that $x(1)(n)$ be taken as initial value[J]. Chinese J of Management Science, 2005, 13(1): 132-134.)

[8] Liu Sifeng, Deng Julong. The rang suitable for GM(1,1)[J]. The J of Grey System, 1999, 11(1): 131-138.

[9] 罗党, 吕健. 几类预测模型模拟精度的比较[J]. 华北水利水电学院学报, 2004, 25(3): 78-80.
(Luo D, Lu J. The comparison with several prediction models in accuracy of simulation[J]. J of North China Institute of Water Conservancy and Hydroelectric Power, 2004, 25(3): 78-80.)

[10] 谢乃明, 刘思峰. 离散GM(1,1)模型与灰色预测模型建模机理[J]. 系统工程理论与实践, 2005, 1(1): 93-99.
(Xie N M, Liu S F. Discrete GM(1,1) and mechanism of grey forecasting model [J]. Systems Engineering-Theory & Practice, 2005, 1(1): 93-99.)

[11] 曾波, 刘思峰, 方志耕, 等. 灰色组合预测模型及其应用[J]. 中国管理科学, 2009, 17(5): 150-155.
(Zeng B, Liu S F, Fang Z G, et al. Grey combined forecast models and its application[J]. Chinese J of Management Science, 2009, 17(5): 150-155.)

[12] 方志耕, 刘思峰, 陆芳, 等. 区间灰数表征与算法改进及其GM(1,1)模型应用研究[J]. 中国工程科学, 2005, 7(2): 57-61.
(Fang Z G, Liu S F, Lu F, et al. Study on improvement of token and arithmetic of interval grey numbers and its GM(1,1) model[J]. Engineering Science2005, 7(2): 57-61.)

[13] Liu S F, Lin Y. On measures of information content of grey numbers[J]. Kybernetes, 2006, 35(5): 899-904.

[14] 邱菀华. 管理决策与应用熵学[M]. 北京: 机械工业出版社, 2001: 158-173.
(Qiu W H. Management decision and application of entropy[M]. Beijing: Machinery Industry Press, 2001: 158-173.)

中国高校科技期刊研究会
加强学术道德和学风建设倡议书

目前, 我国学术研究领域存在的学术风气不正、学术道德腐败等学术不端行为, 对于学术研究、科学发展以及科技期刊自身所产生的负面影响都是不可低估的. 根据教育部关于加强学术道德和学风建设的有关文件精神, 中国高校科技期刊研究会提出如下倡议.

一、加强队伍建设

学术成果发表中学术不端行为的责任主体涉及作者、审者和编辑三个方面. 全国高校科技期刊编辑部首先应重视编辑队伍建设, 加强编辑的文化素养、职业道德教育, 增强编辑的社会责任感, 提倡奉公、敬业的职业精神, 并建立有效的稿件处理规范和制度, 实行编辑问责制.

二、签订出版合同

从合同法的角度看, 作者向期刊投稿的过程和附加说明实际上是一种签订著作权合同的要约. 为避免或减少法律纠纷和不端行为, 编辑部应与作者签订具有约束力的法律文件, 如论文出版合同、版权转让协议或授权书, 也可以要求作者在录用发表前签署诚信声明.

三、完善审稿流程

严格执行"同行评审"制度, 同时结合时代发展的新趋势, 尝试利用网络交互平台实施"公开同行评议".

四、强化检测识别

加强对学术不端行为的检测识别. 应使用有关科技期刊学术不端文献检测系统对来稿进行检测. 要在工作中不断总结、探索行之有效的杜绝学术不端行为的检测识别方法和手段.

五、开展信息交换

编辑部之间应加强信息交换, 利用研究会网站对学术不端者进行内部通报, 建立健全联防机制.

六、处置不端行为

对有学术不端行为者, 可择期发出撤消论文的通告, 并在年度索引中予以删除. 同时, 将其学术不端行为通报给作者单位并上报有关部门.

中国高校科技期刊研究会

2010年11月2日