文章编号: 1001-0920 (2012) 09-1433-04

基于增量学习思想的改进 AdaBoost 建模方法

田慧欣[1], 王安娜[2]

(1. 天津工业大学 a. 电气工程与自动化学院, b. 电工电能新技术天津重点实验室, 天津 300387; 2. 东北大学 信息科学与工程学院, 沈阳 110819)

摘 要: 针对软测量建模的特点以及建模过程中存在的主要问题, 提出了基于 AdaBoost RT 集成学习方法的软测量建模方法, 并根据 AdaBoost RT 算法固有的不足和软测量模型在线更新所面临的困难, 提出了自适应修改阈值 ϕ 和增添增量学习性能的改进方法. 使用该建模方法对宝钢 300 t LF 精炼炉建立钢水温度软测量模型, 并使用实际生产数据对模型进行了检验. 检验结果表明, 该模型具有较好的预测精度, 能够很好地实现在线更新.

关键词: 软测量; AdaBoost; 增量学习; 极限学习机; 精炼炉

中图分类号: TP206 **文献标志码:** A

Improved AdaBoost modeling method based on incremental learning

TIAN Hui-xin[1], *WANG An-na*[2]

(1a. School of Electrical Engineering & Automation, 1b. Key Laboratory of Advanced Electrical Engineering and Energy Technology, Tianjin Polytechnic University, Tianjin 300387, China; 2. College of Information Science and Engineering, Northeastern University, Shenyang 110819, China. Correspondent: TIAN Hui-xin, E-mail: icedewl@163.com)

Abstract: Aiming at the characters and problems of soft sensor, a soft sensor modeling method based on ensemble learning algorithm AdaBoost RT is presented. According to the shortages of AdaBoost RT and the difficulties of updating for soft sensor model, the self-adaptive modified value of ϕ and the incremental learning character added improved methods are proposed. The method is used to establish the soft sensor model of molten steel temperature in 300 t LF. The product data are used to test the model. The test results show that the proposed soft sensor model based on improved AdaBoost RT can improve the prediction accuracy and update real time.

Key words: soft sensor; AdaBoost; incremental learning; extreme learning machine(ELM); ladle furnace(LF)

1 引 言

近年来, 随着人工智能技术的飞速发展, 人工智能方法的软测量技术已在工业生产过程中得到越来越多的应用[1]. 然而, 在实际应用过程中单一的智能方法常常存在一些不足, 制约了实际软测量精度的提高. 集成学习的出现为上述问题的解决提供了一个良好的途径. 集成学习可以有效地提高学习系统的泛化能力, 现今已经成为国际机器学习界的研究热点[2]. 本文将针对软测量建模的特点, 采用适用于回归问题的 AdaBoost RT 集成学习算法建立软测量模型. 另外, 所有的人工智能算法 (包括采用集成算法得到的混合学习机) 为了得到较高的测量精度, 大都要求训练集完整. 但是, 在软测量建模实际应用中, 这一要求却很难得到满足, 这便希望软测量模型能够具有在线更新的能力. 对于传统的更新方法往往使得模型的训练数据不断增加, 即浪费了时间, 又制约了模型的性能; 或只能针对某一特定模型进行更新, 不具有普遍性[3-4]. 这些局限性使得软测量模型的更新成为软测量方法实际应用的瓶颈.

针对上述问题, 本文对原始 AdaBoost RT 集成学习算法进行改进, 提出了自适应修改阈值 ϕ 和增添增量学习性能的改进方法, 使其在克服自身不足的同时, 能够很好地实现在线更新. 将所提出的改进 AdaBoost RT 集成学习算法与神经网络相结合, 并应用于宝钢 300 t LF 炉钢水温度软测量, 取得了较好的预报效果, 而且能够很好地进行在线更新.

收稿日期: 2011-01-08; **修回日期:** 2011-06-27.

基金项目: 国家自然科学基金项目(60843007); 天津市应用基础及前沿技术研究计划项目(11JCYBJC07000).

作者简介: 田慧欣(1978–), 女, 副教授, 博士, 从事复杂工业过程建模、控制与优化的研究; 王安娜(1956–), 女, 教授, 博士生导师, 从事机器学习、智能故障诊断等研究.

2 AdaBoost RT 算法

Boosting 算法最早是为了更好地解决分类问题而提出的, 在分类问题上已被证明是一种有效的方法, 并得到了广泛的应用, 但将其应用于回归问题的研究则起步较晚, 发展得较为缓慢[5-7]. AdaBoost RT 是由 Solomatine 等[8]于 2004 年提出用于解决回归问题的一种算法, 与传统 AdaBoost 算法相比, AdaBoost RT 具有如下特点: 1) 引入了阈值 ϕ, 通过对训练误差和 ϕ 的比较将训练集分成好坏两类; 2) 权值更新参数的计算不同于 AdaBoost R2, 当误差较低时更多地强调学习相对困难的数据; 3) 可以任意定义迭代次数, 没有误差大于 0.5 时不得不停止算法的限制; 4) 最终的输出是通过将弱学习机的输出加权平均得到的. 算法具体描述如下:

算法 1 基本 AdaBoost RT 算法.

Step 1: 输入.

Step 1.1: 建立训练集 $(x_1, y_1), \cdots, (x_m, y_m)$, $y \in R$.

Step 1.2: 确定弱学习算法或弱学习机.

Step 1.3: 确定最大迭代次数.

Step 1.4: 确定判断预报值正确与否的阈值 $\phi, 0 < \phi < 1$.

Step 2: 初始化.

Step 2.1: $t = 1$ 时, 权重分布 $D_t(i) = 1/m$.

Step 2.2: 初始误差 ε_t.

Step 3: 迭代.

Step 3.1: 依据权重分布训练弱学习器.

Step 3.2: 建立回归模型 $f_t(x) \to y$.

Step 3.3: 计算训练集的误差

$$\mathrm{ARE}_t(i) = \left|\frac{f_t(x_i) - y_i}{y_i}\right|. \tag{1}$$

Step 3.4: 计算 $f_t(x)$ 的误差率

$$\varepsilon_t = \sum_{i:\mathrm{ARE}_t(i) > \phi} D_t(i). \tag{2}$$

Step 3.5: 计算 $\beta_t = \varepsilon_t^n, n = 1, 2$ 或 3.

Step 3.6: 更新权重 D_t

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \beta_t, \ \mathrm{ARE}_t(i) \leqslant \phi; \\ 1, \ \text{其他}; \end{cases} \tag{3}$$

这里 Z_t 为标准化因子.

Step 3.7: 令 $t = t + 1$.

Step 4: 输出

$$f_{\mathrm{fin}}(x) = \frac{\sum_{i=1}^{t}\left(\lg\frac{1}{\beta}\right) f_t(x)}{\sum_{i=1}^{t}\left(\lg\frac{1}{\beta}\right)}. \tag{4}$$

3 AdaBoost RT 的改进

分别针对 AdaBoost RT 自身固有的不足以及软测量模型对在线更新的迫切需要, 提出一种自适应修改阈值 ϕ 的改进方法和增添增量学习性能的改进方法. 通过这些改进, 可使 AdaBoost RT 在克服自身不足的同时能够很好地实现在线更新.

3.1 自适应修改阈值 ϕ

AdaBoost RT 算法与 AdaBoost R2 算法相比, 增加了一个需要确定的参数 —— 阈值 ϕ, 它的选择是该算法成功与否的关键因素; 而初始阈值 ϕ 的选择通常通过反复实验获得, 从而增加了整个算法操作的复杂程度, 而且算法的性能对 ϕ 值的选择比较敏感. 如果 ϕ 过小 (例如 $< 1\%$), 则很难获得足够的预测样本; 如果 ϕ 过大, 则不利于噪声的去除以及对困难数据的学习. 这些都将直接影响 AdaBoost RT 性能的优劣, 若不加以处理, 则将影响软测量模型最终的预报性能.

对于阈值 ϕ 的选择在整个 AdaBoost RT 中的重要作用, 本文提出一种自适应修改阈值 ϕ 的方法, 即在整个 AdaBoost RT 迭代学习过程中, 阈值 ϕ 不再是一个固定值, 而是根据每次迭代学习的结果不断变化. 当该次迭代的误差 ε_t 大于前一次迭代的误差 ε_{t-1} 时, 将 ϕ 值增大; 反之, 当该次迭代的误差 ε_t 小于前一次迭代的误差 ε_{t-1} 时, 将 ϕ 值减小. 这样, 可以在整个迭代过程中不断地根据实际情况调节 ϕ 值的大小, 始终向着性能最优的方向寻找最佳阈值, 从而克服了初始阈值选择给 AdaBoost RT 带来的一系列问题.

根据文献 [8] 的研究结果, 可得到下面的结论: 1) 阈值 ϕ 在小于 0.4 的范围内变化时系统的输出比较稳定; 当 ϕ 值大于 0.4 时, 由于过拟合和噪声干扰等因素的影响, 系统对 ϕ 值的变化异常敏感, 整个 AdaBoost RT 集成学习机都处于极度不稳定的状态. 2) 集成学习机的最小误差点通常在 ϕ 值处于 $(0, 0.4)$ 的范围内时得到. 因此, 在采用改进方法之前, 首先将阈值 ϕ 的范围确定在 $0 \sim 0.4$ 之间, 这样既可以简化计算, 又能够避免在寻找最佳阈值时陷入局部极小. 自适应调整阈值的具体操作如下:

1) 确定初始阈值的缺省值为 0.2.

2) 在每次迭代学习 (弱学习器每学习一次) 结束后, 计算本次学习的输出均方根误差

$$e = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2}. \tag{5}$$

3) 阈值的变化率由均方根误差的变化决定, 具体变化过程如下:

$$\begin{cases} \phi_{t+1} = \phi_t \cdot (1 - \lambda), \ e_t < e_{t-1} - \sigma; \\ \phi_{t+1} = \phi_t \cdot (1 + \lambda), \ e_t > e_{t-1} + \sigma; \\ \phi_{t+1} = \phi_t, \ |e_t - e_{t-1}| \leqslant \sigma. \end{cases} \tag{6}$$

其中: σ 是一个自定义的比较小的数; λ 是一个与均方根误差变化率相关的系数

$$\lambda = r \cdot \left| \frac{e_t - e_{t-1}}{e_t} \right|, \tag{7}$$

式中 r 的默认值是 0.5, 也可以由使用者依据各自问题的具体需要确定.

4) 完成本次迭代阈值调整, 继续执行 AdaBoost RT 算法.

通过使用上述自适应修改阈值 ϕ 的方法改进后的 AdaBoost RT, 使用者不必再为确定合适的阈值而耗费大量的时间和精力, 同时还能够保证 AdaBoost RT 的迭代向正确的方向进行, 既可以保证噪声的去除, 又能够保证对困难数据的学习, 最终达到确保 AdaBoost RT 具有较优性能的目的.

3.2　增添增量学习性能

针对现有软测量模型更新方法存在的不足, 将增量学习的思想结合到模型的更新之中, 提出一种为 AdaBoost RT 添加增量学习性能的改进方法. 具体过程如下: 当最后一次迭代完成时, 仍然计算数据集中每个数据的权重, 并将其保留; 当有新数据输入时 (通常需对新数据进行一定的积累), 将原数据集中权重最小的数据剔除, 用新数据代替, 并重新赋予权重; 用该赋予了新权重的新数据集进行 AdaBoost RT 操作, 迭代 1 次或 2 次, 训练获得相应的弱学习机, 并计算它们的误差率; 最后将新的弱学习机与建模时训练好的弱学习机根据它们的误差率集成起来得到新的最终输出, 从而实现对原模型的更新. 该添加增量学习性能的改进方法的具体步骤如下:

算法 2　添加增量学习性能的改进算法.

Step 1 ~ Step 4 与算法 1 相同.

Step 5: 当新数据输入时, 确定新数据累积数目 n, 即每当新数据累积到 n 个时进行以下操作:

Step 5.1: 用新数据代替原数据集中权重最小的数据, 得到新训练数据集

$$S' = [(x'_1, y'_1), \cdots, (x'_n, y'_n), \cdots, (x'_m, y'_m)];$$

Step 5.2: 重新进行权重分布 $D_{T+1}(i) = 1/m$;

Step 5.3: 执行 Step 3, 计算误差率并更新权重, 进行 2 次迭代 (该迭代次数也可由使用者依据不同的具体问题确定), 即 $t = T+1, t = T+2$.

Step 6: 重新计算输出

$$f_{fin}(x) = \frac{\sum_{t=1}^{T+2} \left[\left(\lg \frac{1}{\beta} \right) \cdot f_t(x) \right]}{\sum_{t=1}^{T+2} \left(\lg \frac{1}{\beta} \right)}. \tag{8}$$

图 1 直观地展示了改进 AdaBoost RT 的增量学习更新过程.

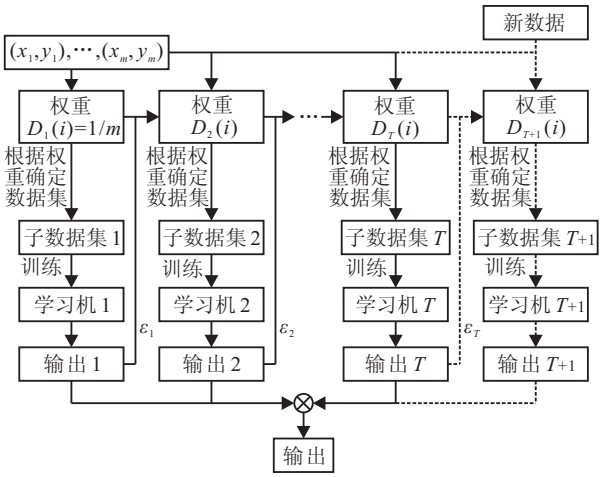


图 1　改进 AdaBoost RT 的增量学习更新过程

与传统的更新方法相比, 采用该方法建立的软测量模型进行更新时, 无须保存历史数据, 可以减少对存储空间的占用; 同时, 该方法在新的训练中利用了历史训练结果, 舍弃了无用样本, 从而显著地减少了后继训练的时间, 并使得学习具有一定的延续性.

4　LF 钢水温度软测量智能建模

4.1　模型结构与算法的确定

针对宝钢 300 t LF 精炼炉的冶炼特点, 采用能量平衡的方法对其进行分析. 确定影响钢水温度的主要因素如下: 冶炼时间, 合金及渣料吸放热, 初始温度, 总的加热电能和钢水重量. 将这 5 个主要影响因素确定为软测量模型的输入, 模型输出为钢水终点温度. 模型中的智能算法选择 ELM, 它是由 Huang 等[9]提出的一种针对单隐含层前馈神经网络的新算法. 研究表明, ELM 具有很好的全局搜索能力和简单易行的特点. 与传统的梯度学习算法 (如 BP 算法) 和 SVM 相比, ELM 的学习速度更快, 更适合在集成学习算法中作为弱学习机. 同时, 使用 ELM 时, 不需要繁琐的确定参数的过程, 这样既可以节省前期准备时间, 又可以很容易地得到合适的参数, 从而提高精度. 最后使用本文提出的改进 AdaBoost RT 对 ELM 进行集成, 得到最终的 LF 钢水温度软测量模型.

4.2　实　　验

取宝钢 300 t LF 精炼炉 2006 年 6 ~ 11 月 380 炉生产数据, 随机抽取 50 炉生产数据作为检验数据. 在剩下的数据中按冶炼时间顺序取前 330 炉生产数据作为训练数据训练模型, 后 30 炉生产数据作为进行模型更新的更新数据. AdaBoost RT 中弱学习机 (ELM) 的个数 (T) 经过实验确定为 8, 新数据的累积数目 $n = 15$. 分别使用未改进和改进后的 AdaBoost RT 建立钢水温度软测量模型, 并采用上述生产数据对模型进行检验.

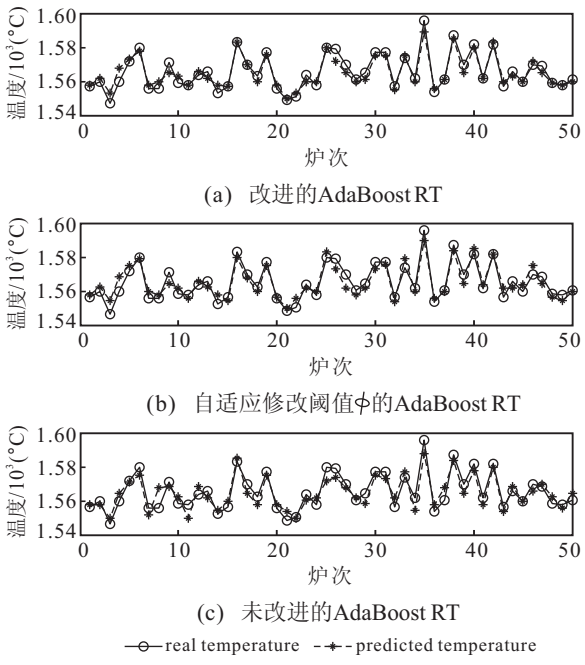
表 1 给出了原始 AdaBoost RT 在不同阈值下的误

差均方根(RMSE)以及使用自适应修改阈值的改进AdaBoost RT的误差均方根. 表1结果表明, 在原始AdaBoost RT集成算法中, 阈值ϕ的大小对AdaBoost RT性能的影响很大, 使用者通常只能通过实验的方法确定一个合适的阈值来保障AdaBoost的性能; 而自适应修改阈值的方法则可以容易地获得AdaBoost RT的最佳性能. 从表1可以得出如下结论: 这种自适应修改阈值的改进方法在克服AdaBoost RT不足的同时, 还可以成功地确保AdaBoost RT拥有最优性能.

表 1 基本 AdaBoost RT 和改进 AdaBoost RT 性能的比较

	阈值ϕ的值	RMSE
AdaBoost RT	0.02	3.330 3
	0.05	3.328 8
	0.08	3.328 0
	0.10	3.329 1
	0.15	3.330 2
	0.20	3.343 3
	0.25	3.361 5
	0.30	3.396 1
	0.35	3.421 5
	0.40	3.765 4
改进的 AdaBoost RT	0.04	3.327 9

检验结果见图2, 分别为采用改进的AdaBoost RT, 单一增加自适应修改阈值ϕ改进方法后的AdaBoost RT和未改进的AdaBoost RT这3种方法建立的钢水温度软测量模型的预报结果. 通过比较可以得出如下结论: 这种新的改进方法可以有效地提高AdaBoost RT的性能, 具有增量学习的更新能力, 可以使基于AdaBoost RT的LF钢水温度软测量模型实现在线更新, 并能够得到更好的预报结果, 从而可以进一步满足实际生产的需求.



(a) 改进的AdaBoost RT

(b) 自适应修改阈值φ的AdaBoost RT

(c) 未改进的AdaBoost RT

图 2 不同建模方法对 LF 钢水温度的预报效果比较

5 结 论

本文针对现有软测量建模方法存在的不足, 采用集成学习算法建立软测量模型, 用以提高软测量的精度. 结合软测量的特点, 针对AdaBoost RT的缺点提出了自适应修改阈值ϕ的方法, 有效地消除了阈值选择对系统的影响. 同时, 通过增加增量学习性能的方法, 使软测量模型具有了在线更新能力. 将所提出的方法应用于LF钢水温度的预报, 使用实际生产数据建立钢水温度软测量模型实验的结果表明, 改进方法有效地改善了AdaBoost RT的性能, 能够确保软测量模型具有增量学习的性能, 进而保证了模型的实时在线更新. 与传统方法相比, 使用所提出的方法建立的软测量模型对钢水温度进行预报, 预报精度有了很大的提高, 完全能够满足实际生产的需要.

参考文献(References)

[1] Fortuna L, Graziani S, Xibilia M G. Soft sensors for product quality monitoring in debutanizer distillation columns[J]. Control Engineering Practice, 2005, 13(4): 499-508.

[2] Dienerich T G. Machine learning research four current directions[J]. AI Magazine, 1997, 18(4): 97-136.

[3] Helland K, Berntsen H E, Borgen O S, et al. Recursive algorithm for partial least squares regression[J]. Chemometrics and Intelligent Laboratory Systems, 1992, 14: 129-137.

[4] Song K, Wang H Q, Li P. Discounted-measurement RPLS algorithm and its application to quality control of rubber mixing process[J]. J of Chemical Industry and Engineering, 2004, 55(6): 942-946.

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

[6] Breiman L. Prediction games and arcing algorithms[J]. Neural Computation, 1997, 11(7): 1493-1518.

[7] Drucker H. Improving regressors using boosting techniques[C]. Proc of the 14th Int Conf on Machine Learning. San Francisco, 1997: 107-115.

[8] Solomatine D P, Shrestha D L. AdaBoost RT: A boosting algorithm for regression prolems[C]. Proc of the Int Joint Conf on Neural Networks. Budapest, 2004: 1163-1168.

[9] Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: A new learning scheme of feedforward neural networks[C]. Proc of Int Joint Conf on Neural Networks. Budapest, 2004: 985-990.

[10] Camdali U, Tunc M. Steady state heat transfer of ladle furnace during steel production process[J]. J of Iron and Steel Research, 2006, 13(3): 18-20.