文章编号: 1001-0920 (2015) 08-1523-04 DOI: 10.13195/j.kzyjc.2014.0941

一种用于软测量建模的增量学习集成算法

田慧欣[a,b], 李 坤[c], 孟 博[a]

(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，c. 管理学院，天津 300387)

摘 要: 针对软测量模型在实际应用中遇到的问题, 结合 AdaBoost 集成学习思想, 提出适用于软测量回归的集成学习算法, 以提高传统软测量模型的精度. 为了克服模型更新技术对软测量实际应用的制约, 将增量学习机制加入软测量集成建模中, 使软测量模型具有在线实时更新的增量学习能力. 对浆纱过程使用新方法建立上浆率软测量模型, 并使用实际生产数据对模型进行检验, 检验结果表明, 该模型具有很好的预测精度, 并能够较好地实现在线更新.

关键词: 软测量；集成建模；增量学习；极限学习机；上浆率

中图分类号: TP206 **文献标志码:** A

An incremental learning ensemble algorithm for soft sensor modeling

TIAN Hui-xin[a,b], *LI Kun*[c], *MENG Bo*[a]

(a. School of Electrical Engineering and Automatic，b. Tianjin Key Laboratory of Advanced Electrical Engineering and Energy Technology，c. School of Management，Tianjin Polytechnic University，Tianjin 300387，China．Correspondent: TIAN Hui-xin，E-mail：icedewl@163.com)

Abstract: Aiming at the characters and problems of the soft sensor, a soft sensor modelling method for the soft sensor regression problem based on the ensemble learning algorithm is proposed to improve the accuracy of the soft sensor. According to the shortages of soft sensor update in practical application, an incremental learning idea is added to the proposed ensemble algorithm for soft sensor modelling. The method is used to establish the soft sensor model of sizing in sizing production. The product data is used to test the model. The results show that the proposed soft sensor model can improve the prediction accuracy and realize online update better.

Keywords: soft sensor；ensemble modelling；incremental learning；extreme learning machine；sizing

0 引 言

近年来, 人工智能技术得到了飞速发展, 基于人工智能技术的各种软测量方法在工业生产过程中也得到广泛的应用. 然而, 在实际应用过程中, 基于单一智能算法的软测量方法往往存在泛化能力有限、容易出现过拟合等不足, 造成软测量模型的精度难以提高, 严重制约了软测量技术在生产过程中的实际运用. 此外, 在实际生产过程中要求软测量模型能够随着生产的进行不断更新, 以确保其测量精度的准确, 这种在线更新能力的优劣直接决定了软测量技术实际应用的有效性. 传统更新方法存在以下问题: 常通过不断增加训练数据对模型进行更新, 使得模型的训练时间不断增加, 若直接减掉部分数据则造成信息缺失; 一些传统更新方法只适用于某一类模型, 不具有应用的广泛性[1-2].

本文针对软测量建模实际特点, 结合 AdaBoost 集成学习算法的思想, 提出适用于软测量回归问题的集成学习算法, 在保证学习连续性的同时有效提高软测量精度. 考虑实际生产对软测量实时更新的需求, 将增量学习思想引入到集成学习中, 提出适用于软测量建模的增量学习集成算法. 该方法具有不断学习新数据的能力, 同时不会遗忘旧知识, 具有在线学习效率高、节省训练时间、所需存储空间小的优点, 为软测量在实际生产过程中的在线应用提供了有效保障.

1 增量学习集成建模方法

集成学习起源于对分类问题的解决, 目前, 其在回归问题上的研究仍少之又少. 1997 年, Freund 等[3]首次将 AdaBoost.M2 延伸到回归问题中, 从而提出了

收稿日期: 2014-06-16；**修回日期:** 2014-09-14.

基金项目: 国家自然科学基金项目(61403277, 61203302)；天津市应用基础与前沿技术研究计划项目(14JCYBJC18900).

作者简介: 田慧欣(1978−), 女, 副教授, 博士, 从事复杂工业过程建模、控制及优化等研究；李坤(1981−), 男, 讲师, 博士, 从事物流优化算法的研究.

AdaBoost.R 集成算法. 同年, Drucker[4]在 AdaBoost.R 的基础上提出了改进算法 AdaBoost.R2, 并验证了其对于回归问题的适用性. 然而, AdaBoost.R2 在迭代时, 误差率一旦大于 0.5, 则必须终止训练, 这成为影响其实际应用的最大缺陷. 2001 年, Polikar 等[5]结合 AdaBoost 思想提出了一种增量集成学习算法——Learn++ 算法, 用于解决多分类问题, 使模型拥有了增量学习的能力[6], 但此类增量学习方法在解决回归问题上的研究仍然处于空白.

针对上述 AdaBoost 算法在回归问题上的不足, 本文在 AdaBoost.R2[4]的基础上增加了误差判定值 e_0, 将误差与误差判定值进行比较, 从而判定学习机的好坏, 对“坏”学习机进行抛弃或重新学习, 克服了 AdaBoost.R2 算法只能在误差小于 0.5 的前提下进行的不足. 同时, 在集成学习过程中加入增量学习思想, 通过权重更新策略的设置, 实现对新数据的增量学习, 即具有对新数据进行学习的能力. 在使用已有弱学习机(映射)时, 若学习效果较差, 则可以生成新的弱学习机(映射)来记录新数据中的信息, 而已有的弱学习机(映射)并不会被舍弃掉, 因此原始数据的信息仍然被保留下来. 该算法通过模型的误差率来计算权重, 并集成最终输出, 在此过程中, 误差率的变化给算法增加了增量学习的性能, 增量学习集成算法描述如下.

从原始数据集中选取 k 个子数据集 S_k, 其中 $k=1,2,\cdots,K$, $S_k=[(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)]$; 确定弱学习机算法; 确定弱学习机的个数 T_k; 误差判定值 e_0.

循环 $k=1,2,\cdots,K$.

初始化 $w_1(i)=D(i)=\dfrac{1}{m(k)}$, $\forall i=1,2,\cdots,m(k)$. 其中: $w_1(i)$ 为样本 i 的权重, $m(k)$ 为数据集的数量.

循环 $t=1,2,\cdots,T_k$.

1) 计算权重

$$D_t(i)=\frac{w_t}{\sum\limits_{i=1}^{m}w_t(i)};$$

2) 根据 D_t 选择训练子集 TR_t 和测试子集 TE_t;

3) 利用 TR_t 训练弱学习机, 得到回归模型 $f_t: x\to y$;

4) 计算 f_i 在 TR_t 和 TE_t 上的误差

$$\mathrm{ARE}_t(i)=\left|\frac{f(x_i)-y_i}{y_i}\right|;$$

5) 计算误差率

$$\varepsilon_t=\sum_{\mathrm{ARE}_t(i)>e_0}D_t(i),$$

若 $\varepsilon_t>e_0$, 则舍弃该模型 f_t, 转到 2);

6) 计算 $\beta_t=\varepsilon_t^n$ ($n=1,2$ 或 3), 并根据 β_t 获得集成回归模型

$$F_k(x)=\frac{\sum\limits_t\left(\lg\dfrac{1}{\beta_t}\right)f_t(x)}{\sum\limits_t\left(\lg\dfrac{1}{\beta_t}\right)};$$

7) 计算 F_t 在训练子集 TR_t 和测试子集 TE_t 上的误差率 $E_t=\sum\limits_{E_t(i)>e_0}D_t(i)$, 若 $E_t>e_0$, 则舍弃该模型 $F_k(k=k-1)$, 转到 2);

8) 计算 $B_K=E_k^n$ ($n=1,2$ 或 3), 根据 B_k 对权重进行更新

$$w_{t+1}=w_t\times\begin{cases}B_t,\ E_t\leqslant e_0;\\ 1,\ \text{else.}\end{cases}$$

根据权重得到最终集成输出

$$F_{\mathrm{fin}}(x)=\frac{\sum\limits_k\left(\lg\dfrac{1}{B_k}\right)F_k(x)}{\sum\limits_k\left(\lg\dfrac{1}{B_k}\right)}.$$

从原始数据集中选取 k 个子数据集 $S_k=[(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)]$, $k=1,2,\cdots,K$, 其中 m 为每个数据集中样本的个数. 确定一种基本智能算法作为集成算法的弱学习机, 并确定需要集成的弱学习机的个数 T_k. 例如, 对于应用最为广泛的 BP 神经网络, 当目标误差设定较大时, 可以将其视为一种弱学习机. 设定误差判定值 e_0. 在训练开始之前, 为每个样本分配一个权重 $D(i)=1/m(k)$, 在后面的训练过程中, 将依据该权重从 S_k 中选取训练子集 TR_t 和测试子集 $\mathrm{TE}_t(S_t=\mathrm{TR}_t+\mathrm{TE}_t)$, 再根据每次循环的训练误差和测试误差对权重进行更新, 对于学习困难的数据, 将增加其被选入训练子集的机会. 测试子集的加入将有助于提高算法的泛化能力.

在对 $t=1,2,\cdots,T_k$ 的每次循环中, 使用由权重 D_t 确定的子训练集 TR_t 对弱学习机进行训练, 从而得到回归模型(映射) $f_t: x\to y$, 这里 $D_t(i)=w_k\Big/\sum\limits_{i=1}^{m}w_k(i)$. 计算训练集 TR_t 和测试子集 TE_t 的每个样本在该映射上的误差 $\mathrm{ARE}_t(i)=\left|\dfrac{f_t(x_i)-y_i}{y_i}\right|$, 在此基础上计算该子数据集 S_k 在映射 f_t 上的误差率 $\varepsilon_t=\sum\limits_{\mathrm{ARE}_t(i)>e_0}D_t(i)$, 并计算 $\beta_t=\varepsilon_t^n$ ($n=1,2$ 或 3). 此时, 若 $\varepsilon_t>e_0$, 则认为该模型未达到设定要求, 舍弃此次训练得到的映射 f_t. 根据下式对样本的权重进行更新

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

其中 Z_t 为标准化因子. 令 $t=t+1$, 使用更新后的权重 $D_{t\mathrm{new}}(i)$ 从 S_k 中重新选取训练集, 并得到新的映射 f_t. 当本次循环结束时, 可以得到 T_k 个映射 $f_1, f_2, \cdots, f_{T_k}$, 此时, 根据每个映射对应的 β_t, 依照式 (4) 对 T_k 个回归模型 (映射) 进行集成, 得到集成回归模型 (映射)

$$F_k(x)=\frac{\sum_{i=1}^{t}\left(\lg\frac{1}{\beta_t}\right)f_t(x)}{\sum_{i=1}^{t}\left(\lg\frac{1}{\beta_t}\right)}. \tag{2}$$

在 $k=1,2,\cdots,K$ 的循环中, 同样计算 TR_t 和 TE_t 中每个样本在集成映射 F_k 上的误差 $E_t(i)=\left|\frac{F_t(x_i)-y_i}{y_i}\right|$ 和误差率 $E_t=\sum_{E_t(i)>e_0}D_t(i)$. 此时, 若 $E_k>e_0$, 则认为该模型未达到设定要求, 舍弃此次训练得到的映射 F_k. 重新选取训练子集并生成新的集成回归模型 (映射) F_k. 计算 $B_K=E_k^n$ ($n=1,2$ 或 3), 并根据 B_k 对每个样本的权重进行更新

$$w_{k+1}(i)=w_k(i)\times\begin{cases}B_t,\ E_t(i)\leqslant e_0;\\ 1,\ \text{else}.\end{cases} \tag{3}$$

更新后的权重用以计算下一次循环时用来确定训练集和测试集的 D_{t+1}. 这种权重更新模式将为算法赋予增量学习的性能. 对于那些信息较容易被回归模型 (映射) 获得的数据, 在下一次的迭代过程中将降低其被选中为训练集的概率, 而信息获取较难的数据被选中为训练集的概率将增加. 算法主要针对那些未能获取到的信息, 随着迭代的进行, 回归模型 (映射) 中的信息量将不断增加, 该算法具有较强的新信息获取能力, 即增量学习能力.

当外循环结束时, 可以得到 K 个映射 $F_1, F_2, \cdots, F_k$, 此时, 根据每个映射对应的 B_k, 依照式 (4) 对 K 个回归模型 (映射) 进行集成, 得到最终的集成回归模型 (映射)

$$F_{\mathrm{fin}}(x)=\frac{\sum_{i=1}^{k}\left(\lg\frac{1}{B_k}\right)F_k(x)}{\sum_{i=1}^{k}\left(\lg\frac{1}{B_k}\right)}. \tag{4}$$

当有新数据需要学习或新信息需要获取时, 只需要增加子数据集 K 的个数, 不需要重复已有的训练, 根据式 (4), 将从新数据得到的映射与已有映射一并集成, 得到最终的映射, 便可实现对整个模型的更新.

增量学习集成算法在拥有增量学习性能的同时, 还继承了 AdaBoost. R 有效提高单一学习机性能的特性. 与 AdaBoost.R 不同的是, 在该算法中, 那些使用已有学习机学习效果较差 (误差较大) 的数据, 被用来生成新的学习机, 以获取数据中蕴含的信息, 并与其他学习机一同生成集成系统. 良好的回归模型学习性能使得该算法能够提供有效的软测量模型, 其增量学习的性能能够保证软测量模型在实际生产中的在线更新能力, 进而提高软测量的实用性.

2　上浆率在线软测量模型

以天纺集团浆纱工艺上浆过程为例, 建立上浆率软测量模型. 针对双浸双压式上浆过程的特点, 将压辊上的纱线和浆槽作为研究对象, 可以认为两者中浆液的总质量在整个浆纱过程中是恒定的. 对于浆槽中浆液的浓度和纱线上浆液的变化进行分析[7], 可以得出影响上浆率的 11 个主要因素: 浆液浓度、浆液粘度、第 1 道和第 2 道压浆辊的压力、浸没辊位置、纱线覆盖系数、纱线张力、浆纱机速度、浆液温度、烘燥温度和采样间隔. 软测量模型的输入为 11 个主要因素, 输出为上浆率. 取天津纺织工程研究院有限公司提供的 490 组上浆过程生产数据建立基于新的增量学习集成算法的软测量模型, 随机抽取 60 组检验模型, 其余数据为生产顺序前 400 组训练模型、后 30 组更新模型. 模型中的子学习机选择学习速度较快的 ELM 极限学习机[8-9]. ELM 的输入层节点为 11, 输出层节点为 1. 模型中的参数通过实验得出 $e_0=0.34$, $T_k=7$. 将 400 组训练数据分为 5 个子数据集 $S_1\sim S_5$, 增量学习集成算法学习的过程如表 1 所示, 根据浆纱生产的需要, 算法的性能用预测精度来衡量, 预测精度由下式计算:

$$P=\frac{N_s}{N_w}\times 100\%. \tag{5}$$

其中: P 为预测精度, N_s 为预测误差 (均方根误差) 小于 0.06 的样本数目, N_w 为所有的测量样本的数目.

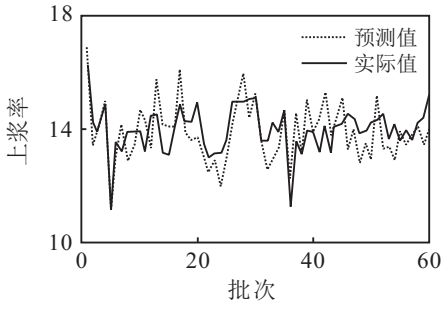
表 1　增量学习集成算法学习过程中精度的变化　%

迭代数	S_1	S_2	S_3	S_4	S_5	更新集	测试集
1	96						86.7
2	96	94					88.3
3	96	95	95				90
4	96	96	97	97			90
5	95	95	97	97	97		93.3
更新	96	96	97	98	97	97	95

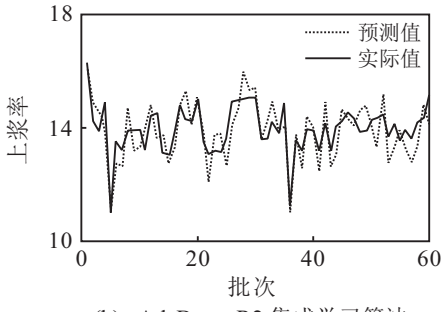
在表 1 中, 测试集为 60 组检验数据. 由于训练数据是按生产过程的时间排列的, 排在后面的数据有可能包含新的信息. 从表 1 中可以看出, 随着数据集的增加, 软测量模型的精度不断提高, 说明新算法具有学习新信息的增量学习能力.

分别采用单一 ELM 智能学习算法、AdaBoost. R2 集成学习算法 (子学习机为 ELM) 和本文提出的增量学习集成算法 (子学习机为 ELM) 3 种方法建立软测量模型, 并对浆纱过程上浆率进行预测, 预测结果

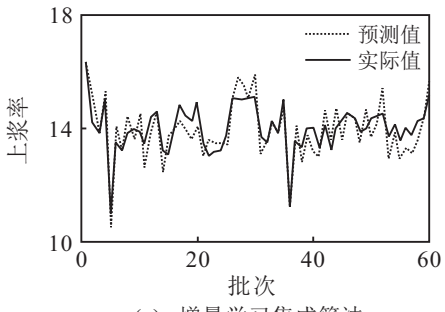
如图1所示.



(a) 单一ELM智能学习算法



(b) AdaBoost.R2集成学习算法



(c) 增量学习集成算法

图 1 软测量模型预测结果比较

3种方法预测误差的比较如图2所示.

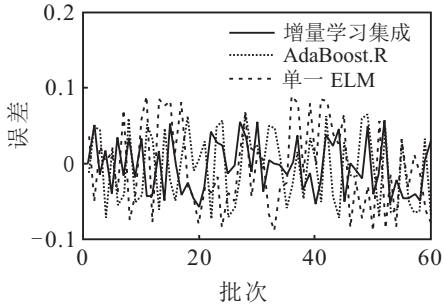


图 2 软测量模型预测误差比较

由图2可以看出, 本文提出的基于增量学习集成算法建立的软测量模型的预测精度最佳, 由于其具有增量学习能力, 可以更好地对新数据中的信息进行学习, 进而可以持续保证模型的预测精度, 实现上浆率预测模型的在线更新, 以满足上浆生产的实际需求.

3 结 论

针对软测量特点和实际应用中存在的问题, 提出了一种新的增量学习集成算法. 该方法能够将从新数据中获取的新信息加入到原有模型中, 使软测量模型能够在确保测量精度的同时, 具有在线更新能力. 利用实际生产数据设计了实验, 实验结果表明, 用于软测量建模的增量学习集成算法不但可以提高模型的预报精度, 还能够很好地实现模型的在线更新, 可以完全满足实际生产需要.

参考文献(References)

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

[2] 胥欣, 江登表, 李勃, 等. 混合高斯模型运动检测算法优化[J]. 计算机应用研究, 2013, 30(7): 2190-2204.
(Xu X, Jiang D B, Li B, et al. Optimization of Gaussian mixture model motion detection algorithm[J]. Application Research of Computers, 2013, 30(7): 2190-2204.)

[3] Freund Y, Schapire R E. 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.

[4] Drucker H. Improving regressor using boosting techniques[C]. Proc of the 13th Annual Conf on Computational Learning Theory. San Francisco, 1997: 208-219.

[5] Polikar R, Udpa L, Udpa S. Learn++: An incremental learning algorithm for supervised neural networks[J]. IEEE Trans on Systems, Man and Cybernetics Part C: Applications and Reviews, 2001, 31(4): 497- 508.

[6] Anita J Patel, Joy S Patel. Ensemble systems and incremental learning[C]. 2013 Int Conf on Intelligent Systems and Signal Proc(ISSP). Gujarat, 2013: 365-368.

[7] 田慧欣, 贾玉凤. 基于集成多支持向量回归融合的上浆率在线软测量方法[J]. 纺织学报, 2014, 35(1): 62-66.
(Tian H X, Jia Y F. Online soft measurement of sizing percentage based on intergrated multiple SVR fusion by Bagging[J]. J of Textile Research, 2014, 35(1): 62-66.)

[8] Huang Guangbin, Zhu Qinyu, Siew Chee. Extreme learning machine: Theory and applications[J]. Neurocomputing, 2006(70): 489-501.

[9] Xiao Dong, Wang Jichun, Mao Zhizhong. The research on the modeling method of batch process based on OS-ELM-RMPLS[J]. Chemometrics and Intelligent Laboratory Systems, 2014, 134(15): 118-122.

(责任编辑: 闫 妍)