

基于 LSSVM 的溶剂油分离建模方法研究

黄燕*, 薄翠梅, 牛超

(南京工业大学电气工程与控制科学学院, 江苏 南京 211816)

摘要:针对传统支持向量机和单一模型建模的缺点,利用某炼油厂溶剂油分离过程中二侧线流量作为建模对象,对最小二乘支持向量机集成学习方法进行了研究。首先利用自适应系数加权模糊(AWFCM)聚类算法对训练样本进行聚类;然后对每一类数据使用最小二乘支持向量机建立子模型,并使用PLS合成函数得到最小二乘支持向量机集成模型;最后通过仿真实验来验证最小二乘支持向量机集成模型预测的精确性。结果表明,该算法在预测精度上有了较大的提高,对过程控制系统中分离效果的预测具有重要指导意义。

关键词:支持向量机;最小二乘法;聚类算法;溶剂油分离;软测量

中图分类号:TP391

文献标志码:A

文章编号:0253-4320(2017)02-0190-04

DOI:10.16606/j.cnki.issn.0253-4320.2017.02.046

Research on modeling methods for solvent oil separation based on LSSVM

HUANG Yan*, BO Cui-mei, NIU Chao

(College of Automation and Electrical Engineering, Nanjing University of Technology, Nanjing 210009, China)

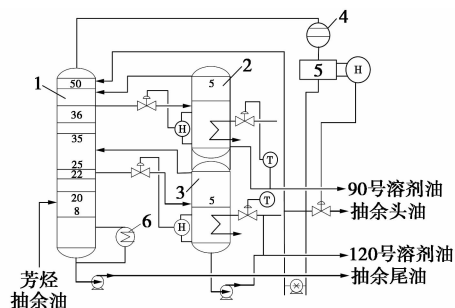
Abstract: As traditional support vector machine and single modeling have some inevitable shortcomings, an integrated learning method of least squares support vector machine (LSSVM) is studied by using two-side line flow in the solvent oil separation of a refinery as the modeling object. The adaptive coefficient weighted fuzzy (AWFCM) clustering algorithm has been used to cluster the training samples. LSSVM is used to establish sub-models for each category of data. The LSSVM integrated models is then obtained by using PLS synthesis function. Finally, simulation experiments are carried out to verify the accuracy of the LS-SVM integrated model. The results show that the proposed algorithm has a great improvement in prediction accuracy and has important guiding significance for the prediction of separation effect in process control system.

Key words: support vector machine; least squares method; clustering algorithm; solvent oil separation; soft measurement

由于溶剂油分离过程的物性数据复杂,许多高度非线性参数会对溶剂油分离过程产生影响,而且该分离过程属于一塔多采^[1-2],采用平衡级的机理建模非常复杂,结果的预测精度差。为了解决机理建模中的这些困难,越来越多的研究人员将软测量建模方法^[3]应用在复杂工业控制过程中,预测控制系统的输出。单一模型的软测量建模方法具有预测结果精度不足,操作参数误差大等问题,并且一些传统的算法也存在着一些不可克服的缺点,例如利用神经网络进行软测量建模时依赖模型的结构,而且在寻找最佳的网络参数时容易陷入局部极值^[4];支持向量机用来解决小样本、非线性及高维数据中有许多优势,当需要处理的数据量大时,就会表现出预测精度不够等缺点^[5-6]。因此研究并利用集成学习的建模方法来解决单一模型带来的缺点。

1 溶剂油分离工艺流程

芳烃抽余油生产溶剂油的分馏塔工艺流程如图1所示。



1—溶剂油分馏塔;2—侧线汽提塔;3—二侧线汽提塔;
4—冷凝器;5—回流罐;6—塔底重沸器

图1 溶剂油分离工艺流程

该系统采用分馏塔装置分离,该塔有50层塔板,设有2个侧线。在生产流程中,非芳烃经泵抽出送到换热器换热后,进入溶剂油分馏塔,并从塔顶部分离出初馏点60℃的馏分,冷却后进入回流罐。一侧线的36层抽出油进入一侧线汽提塔去掉小于60℃的轻组分,从汽提塔底部抽出60~90℃馏分,经分析合格后进入储罐。二侧线的第22层抽出油经汽提塔去掉小于80℃的馏分,得到80~120℃馏

分,称为120号橡胶溶剂油,分析合格进入储罐。与120号溶剂油流量相关的主要影响因素如表1。

表1 与120号溶剂油流量相关的参数变量

变量名称	运行范围	变量名称	运行范围
进料温度/°C	180~220	回流温度/°C	20~35
回流量/(t·h ⁻¹)	5.5~9.5	塔底温度/°C	150~185
一侧线抽出量/(t·h ⁻¹)	1.0~1.5	35层塔板温/°C	90~105
塔顶压力/kPa	1.2~1.7	汽提塔温度/°C	110~125

2 改进的FCM聚类算法

假设 X 是一个包含了 n 个对象 m 维的样本数据集,表示为: $X = \{x_1, x_2, \dots, x_n\}$,其中 $x_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$ 为数据样本 x_i 的特征向量, x_{ij} 是第 i 个对象的第 j 维的属性值,其中 $i = 1, 2, \dots, n, j = 1, 2, \dots, m$ 。

2.1 初始聚类中心的选取方法

由于传统模糊聚类算法中初始聚类质心只能随机选取,容易受到边缘噪声点的干扰^[7],这里采用基于区域密度的初始中心选取方法。该方法中,定义一个密度半径 r ,用于计算在该密度半径中单个数据样本 x_i 的区域密度。设 P_{\min} 是圆区域内样本数的最少个数,即当 x_i 领域内的个数大于 P_{\min} 时,记为有效的高密度区域。首先计算出所有数据样本之间的加权欧式距离,计算公式如下:

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^m W_k (x_{ik} - x_{jk})^2} \quad (1)$$

将计算结果存放在数组 A 中, A_{ij} 表示样本对象 x_i 和 x_j 之间的加权距离,其中($i = 1, 2, \dots, n, j = 2, 3, \dots, n, i \neq j$), n 表示样本的个数。通过对每个样本之间距离与区域半径的比较,将符合条件的样本对象放入数组 Z 中,假设有 k 个对象满足条件。记录下区域密度最高的数据对象 z_i ,将其作为第一个初始化聚类中心 O_1 ,并放入数组 O 中;然后找到满足 $\max(d(z_i, z_j)), j = 1, 2, \dots, k, j \neq i$ 的样本点,将其作为第二个初始聚类中心 O_2 ,并放入 O 中;第三个聚类中心 O_3 根据其 O_1, O_2 的距离的平均值来选择,并且该距离都必须大于半径 r ,即满足 $\max([d_1(O_1, z_i) + d_2(O_2, z_i)]/2) (d_1 > r, d_2 > r)$;以此类推,第 k 个聚类中心 O_k 满足 $\max([d_1(O_1, z_i) + d_1(O_2, z_i) + d_{k-1}(O_{k-2}, z_i)]/(k-1)) (d_{k-1} > r)$,最后得到包含 c 个聚类中心的数组 O 。

2.2 样本属性的系数加权值计算

Fan^[8]在2011年用K-means聚类算法时,对样本对象的各个属性进行了变异系数的计算,并在距

离计算中为每一个维度的属性值加上了权重,实验结果表明,有效地提高了算法聚类的准确率。基于该思想对于数据集 X ,首先计算出所有对象属性的变异系数,变异系数就是对象属性的标准差和平均值的比值,计算公式如下:

$$\beta = \sigma/\mu \quad (2)$$

$$\mu_m = (1/N) \sum_{i=1}^n x_{im} \quad (3)$$

$$\sigma = \sqrt{(1/n) \sum_{i=1}^n (x_{im} - \mu_m)^2} \quad (4)$$

第 i 个对象的第 j 维属性的变异系数值定义为:

$$\beta = \sigma_m/\mu_m = \sqrt{(1/n) \sum_{i=1}^n (x_{im} - \mu_m)^2/\mu_m} \quad (5)$$

定义 $W = \{W_1, W_2, \dots, W_m\}$ 为数据集中样本所有特征属性权值集合,第 j 维的权值定义为:

$$W_j = \beta_j / \sum_{j=1}^m \beta_j \quad (6)$$

带系数加权的数据样本隶属度更新公式为:

$$u_{ij} = \left[\sum_{k=1}^c \left[\left(\sum_{j=1}^m W_j (x_{ij} - c_{kj})^2 \right) / \left(\sum_{j=1}^m W_j (x_{ij} - c_{kj})^2 \right) \right]^{1/\lambda-1} \right]^{-1} \quad (7)$$

式中, c 是数据集聚类的个数; c_{ij} 是第 i 个质心的第 j 个属性值; x_{ij} 是第 i 个样本对象的第 j 个属性值; m 为数据样本特征值的维度; λ 为模糊系数。

质心更新公式为:

$$c_l = \left[\sum_{i=1}^n (u_{il})^\lambda \times x_i \right] / \left[\sum_{i=1}^n (u_{il})^\lambda \right] \quad (8)$$

使用公式(6)、(7)更新数据对象的隶属度及质心的值。准则函数公式如下:

$$J(l) = \left[\left(\sum_{k=1}^c \sum_{i=1}^n u_{ik}^\lambda \times w_i \|v_k - x_i\|^2 \right) / n + (1/c) \|v_i - \bar{v}\|^2 / \left(\min_{i \neq k} \|v_i - v_k\|^2 \right) \right] \quad (9)$$

3 基于AW-FCM的LSSVM集成学习模型的构建

3.1 集成模型合成函数的选择

传统的集成方法一般采用平均系数法或简单加权来集成各个子模型^[9-10]。LSSVM算法建立的各个子模型之间存在着一定的共性,子模型输出间也存在着一定的共性。为了消除输入样本之间共性,这里选择偏最小二乘法(PLS)^[11]作为合成函数。

使用单变量非线性迭代(NIPALS)算法求解PLS合成函数,提取出输入样本的各个主成分,并计算出合成函数的相关系数,函数表示如下^[12]:

$$f_{\text{PLS}}(X) = X^T W (P^T W)^{-1} B q^T \quad (10)$$

式中, X 为输入样本矩阵; W 为 X 的权重矩阵; P 为

X 的负荷矩阵; B 为 PLS 回归系数矩阵。

3.2 LSSVM 集成学习模型的构建

集成 LSSVM 模型的构建流程如图 2 所示, 模型函数可表示为:

$$f(x) = F(s_1(x), \dots, s_c(x)) \quad (11)$$

式中, c 是个体 LSSVM 个数; $s_c(x)$ 表示第 c 个个体子模型。

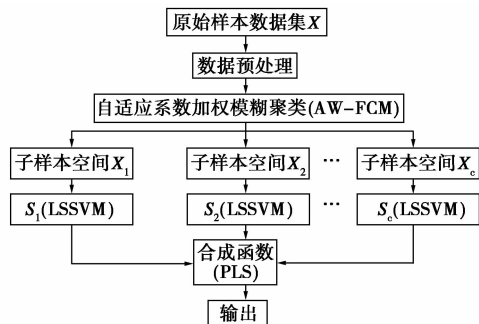


图 2 集成 LSSVM 模型构建流程

4 模型的仿真分析

4.1 集成模型验证

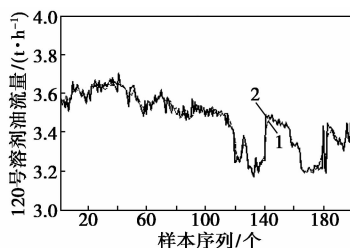
初步筛选出 1 000 组稳态数据, 基于五点三次平滑方法^[13]对数据进行平滑处理以及 PLS 主成分^[14]提取, 选取的主要变量如表 2 所示, 最后将数据分成 2 组, 第一组 800 个数据样本作为训练样本, 第二组 200 个作为测试样本。

首先使用 AWFCM 聚类算法对样本数据聚类。令初始聚类数目 $k=3$, 最低区域密度 $P_{min}=50$, 区域半径的值取所有测试数据样本平均欧式距离的 1.5 倍, 即 $r=1.5R$ 。最大迭代次数 $T_{max}=500$, 允许的误差 $\varepsilon=0.005$, 模糊参数一般的取值在 $(1.5, 2.0)$, 取 $\lambda=1.5$ 。通过 AWFCM 聚类分析, 当 $c=5$ 时准则函数的值最小。对 LSSVM 的惩罚系数 γ 和核参数 σ 使用网格法进行寻优, 使用交叉验证法进行验证。假设 γ 和 σ 的范围是 (γ_1, γ_2) 、 (σ_1, σ_2) , 得到各个子模型的参数如表 2 所示。

表 2 子模型的参数

子模型	S_1	S_2	S_3	S_4	S_5
σ	7.9	6.5	17.6	15.4	5.5
γ	136.5	76.3	213.5	157.6	93.4

对上述得到的各个子模型进行合成得到集成 LSSVM 模型, 并使用该模型对溶剂油分离过程的训练样本和测试样本进行预测, 效果如图 3、图 4 所示。



1—模型预测值; 2—实际测量值

图 3 集成模型训练样本预测结果

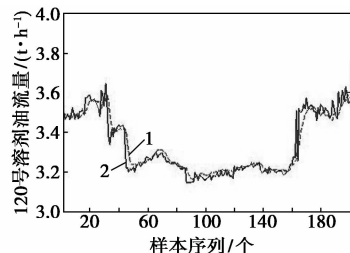


图 4 集成模型测试样本预测结果

模型对训练样本的预测效果比对测试样本的预测效果要好, 通过方均根误差 (RMSE) 以及平均相对误差 (MRE) 来评价模型的性能, 计算公式如下:

$$RMSE = \sqrt{(1/n) \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (12)$$

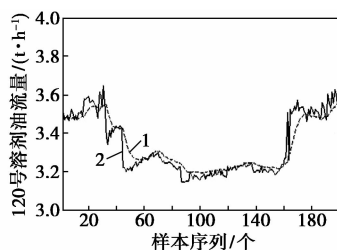
$$MRE = (1/n) \sum_{i=1}^n [|(\hat{y}_i - y_i) / y_i|] \times 100\% \quad (13)$$

通过计算可以得到对训练样本的预测误差 RMSE 为 0.027 t/h, MRE 为 0.96%, 对于测试样本的预测误差 RMSE 为 0.043 t/h, MRE 为 1.03%。而且全部训练样本的预测误差都在 $[-0.02, 0.02]$ 之间, 90% 的预测误差在 $[-0.01, 0.01]$ 之间, 说明模型对训练样本的预测精度非常高; 对于测试样本, 200 个样本中有 121 个测试样本的预测误差在 $[-0.02, 0.02]$ 之间, 占总数的 60.5%, 基本满足对溶剂油分离过程 120 号溶剂油流量的预测。

4.2 多个模型仿真分析

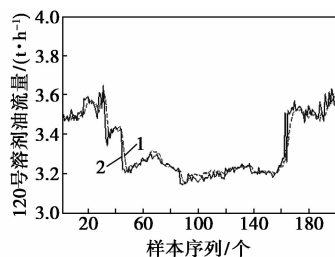
为了验证算法的性能, 针对单个 LSSVM 模型、基于传统模糊 C 均值划分的集成 LSSVM 模型 (FCM-LSSVM-PLS) 以及各模型输出平均值作为合成函数的集成 LSSVM 模型 (AWFCM-LSSVM-A), 在相同的训练集上建立 120 号溶剂油流量软测量模型, 对同样的 200 个测试样本进行预测, 结果如图 5~图 7 所示。

由图 4~图 7 可知, AWFCM-LSSVM-PLS 模型对测试样本的预测效果最好, FCM-LSSVM-PLS 模型的性能略差, 说明对原始数据集聚类优化有助于



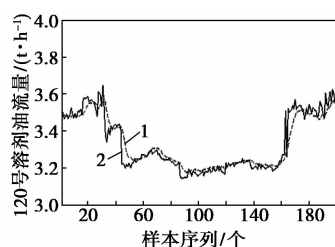
1—模型预测值;2—实际测量值

图5 LSSVM模型预测结果



1—模型预测值;2—实际测量值

图6 FCM-LSSVM-PLS模型预测结果



1—模型预测值;2—实际测量值

图7 AWFCM-LSSVM-A模型预测结果

提高模型的性能;而 AWFCM-LSSVM-A 模型的预测性能相比前面 2 个都差,其预测结果仅仅较好于单个 LSSVM 模型,说明对每个子模型的输出结果简单地求取平均值并不能很好地改善集成模型的预测性能。利用公式(12)、(13)计算出了每个模型对测试样本的 RMSE 和 MRE,结果如表 3 所示。

表3 各个模型的性能分析

算法名称	RMSE/(t·h ⁻¹)	MRE	建模时间/s
LSSVM	0.056	1.37	27
FCM-LSSVM-PLS	0.048	1.18	13
AWFCM-LSSVM-A	0.053	1.32	15
AWFCM-LSSVM-PLS	0.043	1.09	17

从表 3 可知,LSSVM 和 AWFCM-LSSVM-A 模型的预测性能较差, RMSE 超过了 0.05 t/h, MRE 超过了 1.3%, LSSVM 的建模时间最长,进一步说明对每个子模型的输出平均值作为集成模型的输出并不

能很好地体现集成学习的优越性能;而使用 PLS 作为合成策略的 AWFCM-LSSVM-PLS 模型, RMSE 和 MRE 都明显小于 AWFCM-LSSVM-A 模型; FCM-LSSVM-PLS 模型的预测性能也比本文中算法差,说明对初始数据集进行有效地划分能够提高集成模型最终的性能。

4.3 各模型误差分布

各模型的误差分布如表 4 所示。

表4 各模型误差分布

模型	$[-0.05, 0.05]$	> 0.1
AWFCM-LSSVM-PLS	177 (88.5%)	7
AWFCM-LSSVM-A	166 (83%)	14
FCM-LSSVM-PLS	169 (84.5%)	11
LSSVM	161 (80.5%)	16

由表 4 可知,大部分样本分布在 0 ~ 0.05 t/h, AWFCM-LSSVM-PLS 方法在 $[0, 0.05)$ 占的比例最多。

5 结论

基于 LSSVM 集成算法的软测量方法,可以建立有效的复杂工业系统软测量模型。为了提高集成学习方法建模的精度,使用 AWFCM 聚类算法,提高了聚类的精度,使得每个子空间建立的 LSSVM 模型有更好的差异性和代表性。为了更好地消除每个子模型输出之间的共性,使用偏最小二乘法 (PLS) 作为合成函数。

通过对溶剂油分离过程二侧线流量 (120 号溶剂油) 为对象建立了软测量模型,并与传统的单模型以及未改进聚类方法的集成学习模型相比较,实验结果表明,基于 AWFCM-LSSVM 的集成模型有效地提高了对输出结果的预测精度以及对测试样本的泛化能力,能够实现对溶剂油分离过程二侧线流量的快速准确地预测,对 120 号溶剂油的生产有一定的指导意义。

参考文献

- [1] 范景丽. 普通溶剂油市场行情回顾及展望[J]. 辽宁化工, 2012, 41(8): 785-788.
- [2] 吴文波. 基于多生产方案溶剂油生产装置的建模与应用[D]. 上海: 华东理工大学, 2011.
- [3] 曹鹏飞, 罗雄麟. 化工过程软测量建模方法研究进展[J]. 化工学报, 2013, 64(3): 788-800.
- [4] 吕游. 基于过程数据的建模方法研究及应用[D]. 保定: 华北电力大学, 2014.

传感器。PID 气体传感器是一种通用式气体传感器,他对大多数挥发性有机气体都会有较强的响应信号^[8]。检测原理是在高电压的作用下,通过紫外光源的照射使被测物质发生电离,通过采集极板进行检测,原理结构如图 1 所示。

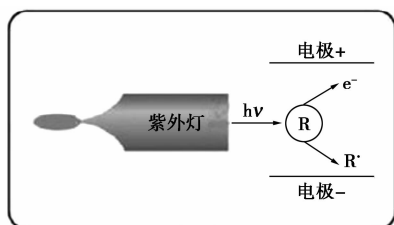


图 1 PID 检测原理图

2 色谱柱的制备

芯片式微型气相色谱柱是在硼硅玻璃上刻蚀出横截面为圆形的连通式沟道,通过键合技术制作而成。芯片的整体布局与结构如图 2 所示。

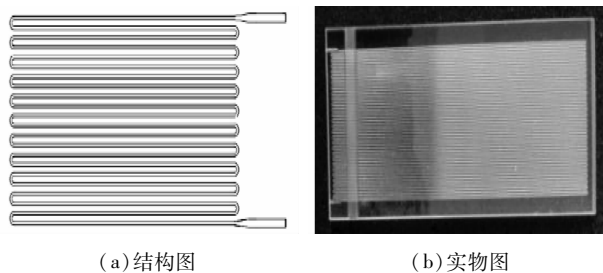


图 2 色谱柱图

沟道的布局采用 S 形往复分布,这种分布的好处在于可以使气体在气路中流动均匀。沟道的形状如图 3 所示。在设计好通道形状和结构布局后制作

掩模板,利用掩模板保护需要键合的非刻蚀区域^[9],加工 2 块同等规格的玻璃芯片。通过激光定位对准刻蚀后的通道,采用热键合工艺对玻璃芯片进行键合,完成键合过程^[10]。刻蚀与键合过程如图 4 所示。

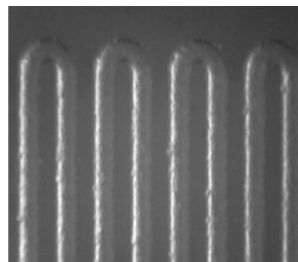


图 3 显微镜下通道照片

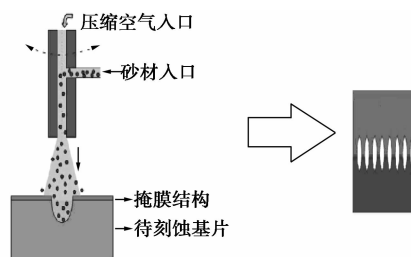


图 4 微型色谱柱制造原理图

色谱柱制作完成后,进行固定相的选择,并填充涂敷固定相,文中使用了 3 组检测模块,针对不同的模块所选择的固定相是不同的。在固定相的使用上选用了 3 种不同的物质组合,分别为:CB-40(40% 苯基 + 60% 二甲基聚硅氧烷)、CB-60(60% 苯基 + 40% 二甲基聚硅氧烷)、OV-1701(14% 氰丙基苯基 + 86% 二甲基聚硅氧烷),将 3 种不同的固定相

(上接第 193 页)

[5] 丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报,2011,40(1):2-10.

[6] 周开乐. 模糊 C 均值聚类及其有效性检验与应用研究[D]. 合肥:合肥工业大学,2014.

[7] 唐苦. 基于聚类的多模型动态软测量建模方法[D]. 上海:华东理工大学,2014.

[8] Fan A. K-means clustering algorithm based on coefficient of variation[J]. Computer Engineering & Applications,2012,4(35):2076-2079.

[9] Zhao Feng, Liu Hanqiang, Jiao Licheng. Spectral clustering with fuzzy similarity measure[J]. Digital Signal Processing, 2011, 21(6):701-709.

[10] Michela Ott, Francesca Pozzi. Digital games as creativity enablers for children[J]. Behaviour & Information Technology, 2012, 31

(10):1011-1019.

[11] 何开锋,钱炜祺,张勇. 偏最小二乘法在气动数据建模中的应用[J]. 宇航学报,2014,35(3):277-282.

[12] Galicia H J, Peter He Q, Wang J. Comparison of the performance of a reduced-order dynamic PLS soft sensor with different updating schemes for digester control[J]. Control Engineering Practice, 2012,20(8):747-760.

[13] Basu B, Fofoula-Georgiou E, Sharma A S. Chaotic behavior in the flow along a wedge modeled by the Blasius equation[J]. Nonlinear Processes in Geophysics,2011,18(2):171-178.

[14] Nascimento P A M, Carvalho L C D, Júnior L C C, et al. Robust PLS models for soluble solids content and firmness determination in low chilling peach using near-infrared spectroscopy (NIR)[J]. Postharvest Biology & Technology,2016,111:345-351. ■