基于核函数及参数优化的 KPLS 质量预测研究*

陈 路,郑 丹,童楚东
(宁波大学 信息科学与工程学院,浙江 宁波 315211)

摘 要:核偏最小二乘(KPLS)在工业过程监测和质量预测中得到了广泛的应用,核函数和核参数的选取对 KPLS 质量预测结果有重要影响。然而,如何选择核函数类型和核参数一直是该方法应用的瓶颈。针对以上问题,提出一种改进遗传算法的核函数优化方法。该方法将核的种类及核参数作为优化的决策变量,以均方根误差为目标,分别从编码方案、遗传策略、适应度函数优化、交叉和变异算法等方面进行设计,以保证核函数种类的多样性,利用 2 折交叉验证法对训练结果进行验证。以田纳西-伊斯曼过程(TE)与 MATLAB 结合进行仿真实验,仿真结果表明,该方法能寻找到最优核函数以及其核参数,具有很好的稳定性和一致性。

关键词:核偏最小二乘;遗传算法;质量预测;k 折交叉验证
中图分类号:TN081;TP277 **文献标识码**:A DOI:10.16157/j.issn.0258-7998.201259

中文引用格式:陈路,郑丹,童楚东. 基于核函数及参数优化的 KPLS 质量预测研究[J].电子技术应用,2021,47(12):100-104.
英文引用格式:Chen Lu,Zheng Dan,Tong Chudong. The optimization of the kind and parameters of kernel function in KPLS for quality prediction[J]. Application of Electronic Technique,2021,47(12):100-104.

The optimization of the kind and parameters of kernel function in KPLS for quality prediction

Chen Lu,Zheng Dan,Tong Chudong
(Faculty of Electrical Engineering and Computer Science,Ningbo University,Ningbo 315211,China)

Abstract:Kernel partial least squares(KPLS) has been widely used in industrial process monitoring and quality prediction. The choice of kernel function and kernel parameters has an important impact on the KPLS quality prediction results. However, how to choose the kernel function type and kernel parameters has always been the bottleneck of the application of this method. To solve the above problems, a kernel function optimization method based on improved genetic algorithm is proposed. In this method, the kernel type and kernel parameters are used as the optimal decision variables, and the root mean square error is targeted. It is designed in terms of coding scheme, genetic strategy, fitness function optimization, crossover and mutation algorithms to ensure the variety of kernel functions, and uses the 2-fold cross-validation method to verify the training results. The Tennessee-Eastman Process(TE) is combined with MATLAB for simulation experiments. The simulation results show that the method can find the optimal kernel function and its kernel parameters, and has good stability and consistency.
Key words:kernel partial least squares;genetic algorithm;quality prediction;k-fold cross-validation

0 引言

质量预测与分析是实现工业过程闭环控制的基础和关键[1]。基于 KPLS 的方法可以提高质量预测精度,许多研究人员以 KPLS 方法为基石,提出了许多解决非线性问题的方法[1-8]。

核函数是 KPLS 方法的关键,而 KPLS 选择核函数并不是任意的,必须要满足 Mercer 定理。特定的内核函数选择隐含地决定了映射和特征空间。在 KPLS 中,由于提取系统非线性特征的程度是基于核函数的,因此核函数的选择是最重要的。如何给基于 KPLS 的质量预测选择理想的核函数和核参数是一个开放的问题[9-10]。而且,一旦设置了核函数,就需要设置适当的核参数。但是,没有一个理论框架能寻找到指定核函数的参数最最优值,也就是说基于 KPLS 的质量预测很大程度上取决于选择的核函数和核参数。

目前,关于如何选择核函数的种类和参数来进行质量预测的研究还没有报道。在过去的几年里,Huang 提出了一种用于特征选择和参数优化的遗传算法[11]。Adriano 将此方法应用于软件工作量估算[12]。Jia 提出了一种改进遗传算法,将改进的遗传算法用于工业过程故

*基金项目:国家自然科学基金项目(61773225);浙江省自然科学基金项目(LY20F030004)

障检测，从而提升了故障监测率[13]。此外，Liu 提出一种基于多核线性学习的 KPLS 方法，该方法使用自适应遗传算法来选择核函数的参数和权重[14]。这些研究多数都是固定一种核函数，然后对其参数进行优化，或者通过改进遗传算法提高优化速度。但是，核函数种类不止一种，只选择其中一种核函数具有局限性。基于这些研究成果，本文结合 KPLS 质量预测，建立同时优化核函数及参数的优化模型 GA-KPLS-V。通过组合编码方式，将不同核函数及参数类型结合在一起，在该模型中，核函数类型和核参数被视为决策变量，将质量预测结果 RMSE 作为目标，并采用改进遗传算法对模型进行求解。仿真结果表明，该方法不仅能选出最优核函数，还能选出最优核参数。

1 基本方法介绍

1.1 KPLS 概述

KPLS 算法的主要思想是通过一个非线性映射将输入数据 $x_i \in R^m$ 映射到一个高位的特征空间 H，特征空间的维数可以任意大甚至无穷大，然后在高位特征空间 H 中构建线性 PLS 模型。由于特征空间 H 的维数很高，不可能直接计算出得分向量、权值向量和回归数值，因此必须对原始空间的运算公式进行变换，使它只包含映射后数据的内积运算，而内积运算可以由原始空间定义的核函数来表示，即：

$$\boldsymbol{K}(x_i,x_j)=\langle \Phi(x_i),\Phi(x_j)\rangle \tag{1}$$

其中，$\boldsymbol{K}$ 为 $n\times n$ 维核矩阵，表示非线性映射所选择的核函数；$\Phi(x_i)$为映射后的输入数据。代表性的核函数如下：

(1)径向基核函数：

$$k(\boldsymbol{x},\boldsymbol{y})=\mathrm{e}^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{c}} \tag{2}$$

(2)Sigmoid 核函数：

$$k(\boldsymbol{x},\boldsymbol{y})=\tanh(\beta_0\langle \boldsymbol{x},\boldsymbol{y}\rangle+\beta_1) \tag{3}$$

(3)多项式核函数：

$$k(\boldsymbol{x},\boldsymbol{y})=(1+\langle \boldsymbol{x},\boldsymbol{y}\rangle)^d \tag{4}$$

其中，c、d、β_0、β_1 是核函数的参数，使用者依据经验决定。依据经验选择核函数和核参数进行质量预测会造成预测结果不稳定，并且使用不同核函数的最优预测结果也不相同。同时，不同核函数的参数类型和个数也不相同，这也给同时优化多种核函数和核参数增加了难度。径向基核函数和多项式核函数一直满足 Mercer 定理，但是 Sigmoid 核函数只有在特定的和下才满足条件[14]。

1.2 遗传算法基本原理

遗传算法是把问题的解决方案用某种编码方式表示，产生初始种群并根据适应度函数计算适应度。然后，再利用选择、交叉和变异操作，不断迭代优化，直到找到最优解。传统的遗传算法编码方式一般是相同的，不然无法进行选择、交叉、变异操作。本文采取的是同时优化多种核函数及核参数，因此，编码、适应度函数、选择、交叉、变异等操作都需要重新设计。

2 基于 KPLS 的质量预测优化方法

2.1 优化指标

在质量预测方面，其精度是衡量其算法对数据预测是否准确的一大重要指标，用于衡量预测值对其真实值偏离程度[15]。其计算或表示方法在不同的领域会有些许的不同，核心公式是：

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \tag{5}$$

式中，y_i 表示实际值，$\hat{y}_i$ 表示预测值，n 表示全部样本数。RMSE 称作均方根误差，在训练集中的均方根误差表示为 RMSEC，在测试集中的均方根误差表示为 RMSET。本文将采用均方根误差 RMSEC 作为确定最优的优化指标。

遗传算法在进化搜索过程中是以目标函数(即适应度函数)为依据来引导寻优过程的。本文的优化目标是质量预测精度均方根误差 RMSE，因此，适应度函数可以设计为：

$$\mathrm{fitness}=\mathrm{RMSEC} \tag{6}$$

本文问题为确定 KPLS 中核函数类型及参数，为保证核函数种类的多样性，避免某种核函数种群过早消失，各类型子类应保证一定的数量[13]。因此，为了满足要求，需要重新构造适应度函数。

$$\mathrm{fitness\_}t=\frac{\mathrm{fitness}}{f\_s_i}\times\frac{g-g_i}{(k-1)g}\times w+(1-w)\times\frac{\mathrm{fitness}}{f\_s} \tag{7}$$

式中，$w=\mathrm{e}^{-m+1}$ 为加权因子，m 为遗传代数，f_s_i 为某一代中第 i 类核函数个体适应度总和，f_s 为某一代中所有个体适应度总和。

2.2 遗传算法设计

2.2.1 编码

遗传算法常见的编码方式包括二进制编码、浮点数编码、格雷码和符号编码等。基于 KPLS 质量预测中需要寻优的参数包括核函数的种类及其核参数，种类不同的数其参数个数、范围也不相同。针对本文的问题，采用了混合编码策略。如图 1 所示，染色体有三部分：第一部分为核函数类型，采用二进制编码；其余部分为核参数采用浮点数编码。当选取径向基核函数和多项式核函数时，第二部分有效，第三部分无效。当选取 Sigmoid 核函数时，第二、第三部分均有效。

图 1 编码结构

2.2.2 初始种群设置

由于核函数种类少，因此，核函数种类采用枚举方法，设核函数种类为 v。而相对应的核参数范围较大，

不能采用枚举的方法。设子群个体数为 g_i，则总群个体数为：

$$g=\sum_{i=1}^{v} g_i \tag{8}$$

当确定种群大小后，采用随机办法产生种群中的个体，并令初始子种群个数为：

$$g_i=\frac{g}{v} \tag{9}$$

2.2.3 遗传操作

本文选择轮盘赌方法进行个体选择。由于采用的是组合编码，因此交叉操作比一般问题复杂。如果两个父本核函数种类相同，直接进行一般交叉操作；若种类不同，分别在父代中寻找相同个体进行交叉，如果没有找到相同个体，则随机产生一个个体交叉。变异就是以很小的变异概率 p_m 随机地改变种群中个体的某些基因值，由于采用的组合编码，因此变异位只选择参数位变异。

2.3 k 折交叉验证

在使用不同核函数建立 KPLS 质量预测模型时，可能会出现使用高斯核函数的 KPLS 质量预测模型在训练集上的学习能力好于使用其他核函数的 KPLS 质量预测模型，但是在测试集上的表现却比使用其他两种核函数的 KPLS 质量预测模型差。这种现象在机器学习中称为“过拟合”。

采用 k 折交叉验证法(K-Fold Cross Validation)能够有效避免以上情况。首先，将训练集划分为 k 个大小相同的互斥子集。然后，依次选取不同的区作为验证集，其余的 $k-1$ 个区作为训练集，每次验证都会得到一个均方根误差，最后总误差为所有误差的平均值，如图 2 所示。

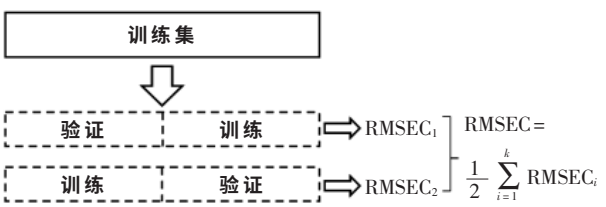

图 2 k 折交叉验证原理图

k折交叉验证法评估结果的稳定性在很大程度上取决于 k 的取值，一般情况下 k 的取值区间为[2,10]。本文采用 2 折交叉验证法进行评估。k 值越大，偏差会越小而复杂度会越大，本文研究重点为找到最佳核函数以及核参数，而 k 在取值范围内进行交叉验证得到的最佳核函数和核参数相同。

2.4 KPLS 优化过程

KPLS 优化过程如下：

(1)将采集的正常运行数据划分为训练集和测试集，对训练集数据进行标准化处理，保存均值和方差。

(2)确定核函数的种类、初始种群大小和终止条件。根据核函数参数的取值范围，随机生成核函数参数并进行编码。

(3) 通过 k折交叉验证对种群中的每个体进行 KPLS 建模并计算适应度值。

(4)进行遗传操作：选择、交叉、变异，进入下一代。

(5)判断是否满足终止条件，如果不满足，重复步骤(3)~(4)；否则，停止迭代。

基于核函数及参数优化的 KPLS 质量预测基本算法示意图如图 3 所示。

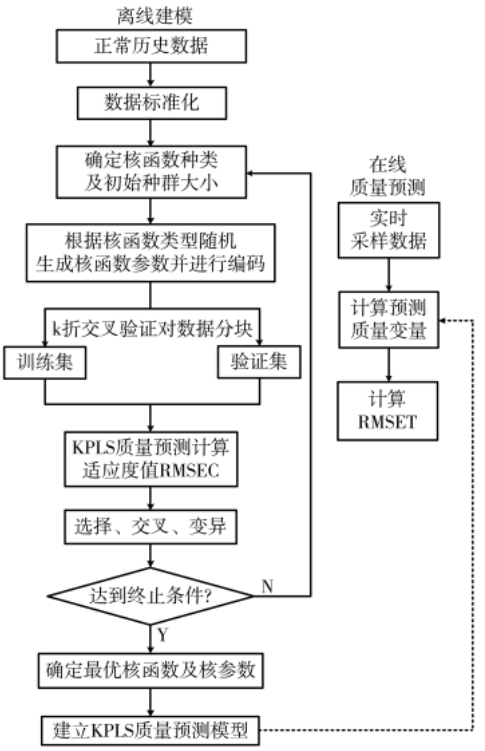

图 3 KPLS 质量预测优化过程流程图

3 仿真实验

Tennessee Eastman(TE)过程是基于实际工业过程的仿真实例，它由美国 Tennessee Eastman 化学公司过程控制部门的 Downs 和 Vogel 于 1993 年提出。此过程已经被广泛作为连续过程的策略、监视、诊断的优化的研究平台。此过程有 4 种反应物(A、C、D、E)，生成两种产物(G 和 H)。此外，还包含一种惰性物质 B 及副产物 F。该过程有 22 个连续的过程测量值、12 个操作变量和 19 个混合测量值。在仿真过程中采样数据的时间间隔是 3 min。本实验共选取 22 个连续变量核 11 个操作变量作为输入变量 X，同时选取物流 9 中组分 G 的摩尔含量作为输出变量 y。首先在正常状况下让仿真程序运行 25 h 得到 960 个正常样本，然后将前 760 个正常样本作为训练集数据，将后 200 个正常样本作为测试集数据。

针对径向基核函数、多项式核函数、Sigmoid 核函数 3 种核函数进行寻优。其中，径向基核函数的参数取在(0.1,50)范围内；多项式核函数的参数 d 取为 1~8 的正整数；Sigmoid 核函数的参数 β_0 和 β_1 都取为(1,8)。设初

始种群个子类个体数为 8，总种群数为 24，根据核函数的数的定义域随机产生初始种群，遗传代数为 20。

本文称随机选取高斯核函数参数的 KPLS 质量预测方法为 KPLS-G，随机选取多项式核函数参数的 KPLS 质量预测方法为 KPLS-P，随机选取 Sigmoid 核函数参数的 KPLS 质量预测方法为 KPLS-S，遗传算法优化选取高斯核函数参数的 KPLS 质量预测方法为 GA-KPLS-G。

在训练集和测试集相同的前提下，分别使用 KPLS-G、KPLS-P、KPLS-S、GA-KPLS-G、GA-KPLS-V 进行 100 次质量预测，预测结果如图 4 和图 5 所示，优化过程如图 6~图 9 所示。

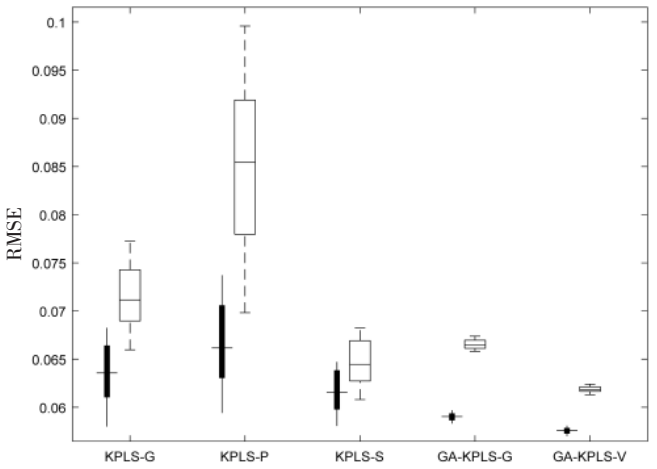

图 4 5 种不同核函数选取方法预测结果对比(变量 35)

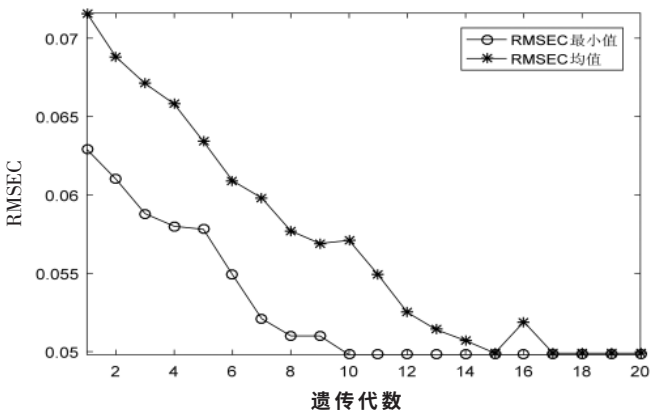

图 5 均方根误差(变量 35)

对于变量 35，从图 4 可以看出，其中实心箱体为训练结果，空心箱体为测试结果。基于 KPLS-G、KPLS-P和 KPLS-S 的质量预测结果数据相对分散，说明随机选取核函数参数会导致预测结果不稳定。基于 GA-KPLS-G 的质量预测结果数据相对集中，说明预测结果具有较好的稳定性和一致性。但是，核函数种类不止一种，选择常用的高斯核函数进行优化建立质量预测模型只能得到基于高斯核函数的最优解，从图中可以看出 KPLS-S 的预测精度是有可能高于 GA-KPLS-G 的预测精度的。本

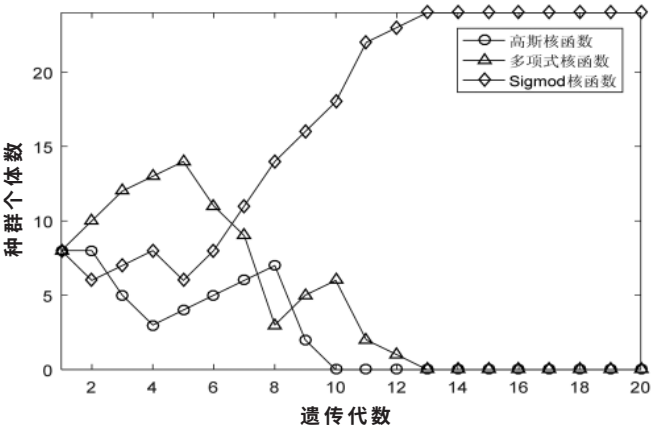

图 6 核函数种群变化(变量 35)

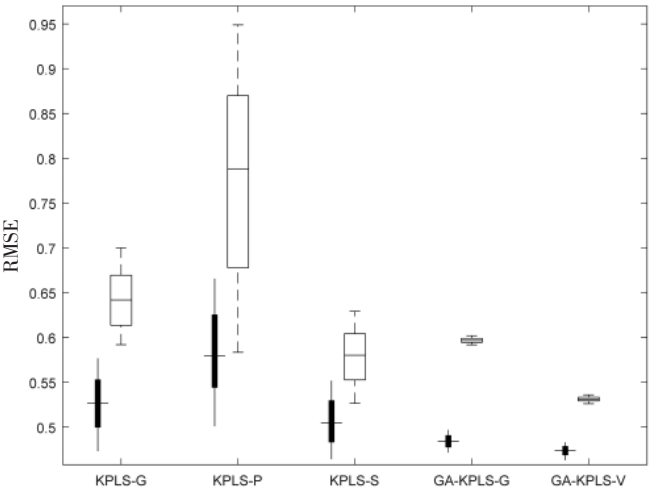

图 7 5 种不同核函数选取方法预测结果对比(变量 40)

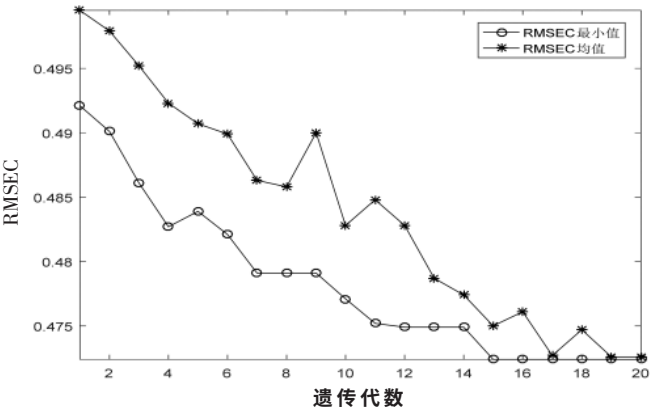

图 8 均方根误差(变量 40)

文提出的 GA-KPLS-V 质量预测结果的中位数最低且预测结果相对集中，说明其平均质量预测精度最高且具有较好的稳定性和一致性。其中，最优核函数为 Sigmoid 核函数，最优核参数 β_0=1.27，β_1=7.81。

图 6 为遗传过程中 RMSEC 值的变化曲线，其中“○”为最小值，“*”为平均值。可以看出，随着遗传代数增加，RMSEC 均值和最小值逐渐降低并趋于平稳。图 7 为

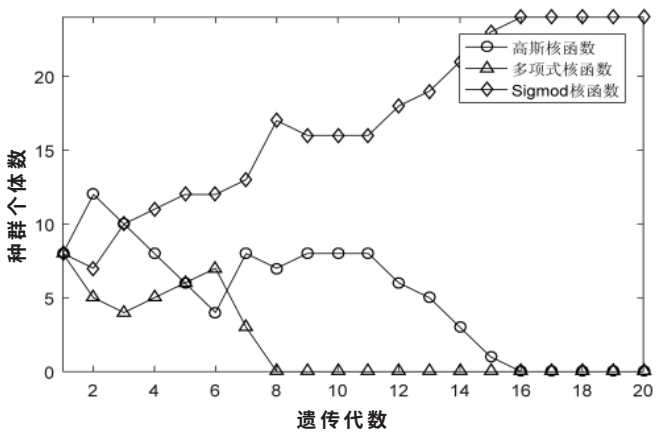

图 9 核函数种群变化(变量 40)

遗传过程中种群变化趋势,可以看出,随着遗传代数的增加,高斯核和多项式核的种群逐渐减少最终为 0,而 Sigmoid 核种群逐渐增多,最终达到最大个体数。

对于变量 40,从图 7 可以看出,本文所提出的 GA-KPLS-V 质量预测结果依旧是 3 种方法中效果最好的。其中,最优核函数为 Sigmoid 核函数,最优核参数 β_0=0.22,β_1=3.41。

图 8 可以看出,随着遗传代数增加,RMSEC 均值和最小值逐渐降低并趋于平稳。从图 9 可以看出,随着遗传代数的增加,Sigmoid 核种群逐渐增多,最终达到最大个体数。常用核函数类型不止一种,选择一种核函数进行参数优化具有局限性。选择优化多种核函数及参数的质量预测方法,能够同时选择最优核函数及核参数,对复杂过程质量预测更稳定和准确。

4 结论

本文在 KPLS 的应用中,针对基于 KPLS 在线质量估计和预测中核函数和核参数难以确定的问题,提出了优化方法。该方法以遗传算法为优化算法,以核函数的种类和参数为优化决策变量,以均方根误差 RMSE 为优化目标。该方法的优势在于,遗传算法的解收敛于全局最优解,同时,重新设计适应度函数和交叉变异算法,保证了核函数种群的多样性。仿真结果表明,该方法能选出最优核函数和和参数,且提高了模型预测精度。

参考文献

[1] 刘春燕,于春梅,闫广峰.一种基于特征子空间的改进动态核主元分析方法[J].计算机应用研究,2016,33(12):3713-3716.

[2] 张敏,程文明.一种基于局部模型的多工况过程质量预测方法[J].计算机应用研究,2014,31(6):1740-1743.

[3] 姜庆超,颜学峰.基于局部-整体相关特征的多单元化工过程分层监测[J].自动化学报,2020,46(9):1770-1782.

[4] 汪司飞,黄菲.基于 K 均值聚类的 KPCA 在故障诊断中的应用[J].计算机应用与软件,2013,30(4):120-123.

[5] CHEW W,SHARRATT P.Trends in process analytical technology[J].Analytical Methods,2010,2(10):1412-1438.

[6] 张绪红,肖应旺.基于多向核偏最小二乘的间歇过程在线监控[J].计算机与应用化学,2017,34(6):434-440.

[7] 彭开香,马良,张凯.复杂工业过程质量相关的故障检测与诊断技术综述[J].自动化学报,2017,43(3):349-365.

[8] 李征,王普,高学金,等.基于信息增量矩阵的多阶段间歇过程质量预测[J].化工学报,2018,69(12):245-253.

[9] BENNETT K P,EMBRECHTS M J.An optimization perspective on kernel partial least squares regression[J].Nato Science Series Sub Series III Computer and Systems Sciences,2003,190:227-250.

[10] DANIEL J,HERNANDEZ A.KPLS optimization approach using genetic algorithms[J].Procedia Computer Science,2020,170:1153-1160.

[11] HUANG C L,WANG C J.A GA-based feature selection and parameters optimization for support vector machines[J].Expert Systems with Applications,2006,31(2):231-240.

[12] ADRIANO L I O,PETRONIO L B,RICARDO M F L,et al.GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation[J].Information and Software Technology,2010,52(11):1155-1166.

[13] Jia Mingxing,Xu Hengyuan,Liu Xiaofei,et al.The optimization of the kind and parameters of kernel function in KPCA for process monitoring[J].Computers and Chemical Engineering,2012,46:94-104.

[14] Liu Shaowei,Tang Jian,Yan Dong.Multi-Kernel partial least squares regression based on adaptive genetic algorithm[C]//2015 International Conference on Automation,Mechanical Control and Computational Engineering.Guilin:Atlantis Press,2015.

[15] Zhang Yingwei,Teng Yongdong.Process data modeling using modified kernel partial least squares[J].Chemical Engineering Science,2010,65(24):6353-6361.

(收稿日期:2020-12-29)

作者简介:

陈路(1993-),男,硕士,主要研究方向:数据驱动的工业过程监测。

郑丹(1994-),女,硕士,主要研究方向:数据驱动的工业过程监测。

童楚东(1989-),男,博士,教授,主要研究方向:数据驱动的工业过程监测。

扫码下载电子文档

版权声明