

研究论文

微粒群算法优化化工建模训练集

张运陶, 高世博

(西华师范大学应用化学研究所, 四川 南充 637002)

摘要: 提出两种均以微粒群 (PSO) 算法对原始训练集随机抽样优化, 再结合机器学习算法建立预测模型的 PSO 算法优化化工建模训练集的思路。思路 1 首先以模型交叉验证的均方误差函数 mse 最小为目标优化训练集, 再通过对验证集预测, 从平行运行得到的多个优化训练集中确定最优训练集用于建模。思路 2 借鉴提高 BP 神经网络泛化能力的初期终止 (early stop) 法, 以对验证集预测的 mse 最小为目标优化训练集, 再通过对测试集预测, 从平行运行得到的多个优化训练集中确定最优训练集用于建模。通过仿真实验研究和对某炼油厂调和汽油生产数据的具体分析应用, 表明本文思路可以较大幅度提高模型的预测准确性, 在化工建模中具有推广应用价值。

关键词: 微粒群算法; 优化训练集; 仿真; 化工建模

中图分类号: TP 18; TE 6

文献标识码: A

文章编号: 0438-1157 (2008) 04-0964-06

Particle swarm optimization optimizing training set in chemical engineering modeling

ZHANG Yuntao, GAO Shibo

(Institute of Applied Chemistry, China West Normal University, Nanchong 637002, Sichuan, China)

Abstract: Two methods were proposed to optimize the training set of chemical engineering modeling based on particle swarm optimization (PSO) with random sampling. The forecasting model was formed by integrating PSO and other machine learning arithmetic which was used to model. In method 1 the optimized training set was acquired firstly by optimizing the training set which set the goal of minimizing the mean of squared errors (mse) in cross validation, secondly by making the prediction to the validation set, and lastly by choosing the one that gave the best prediction results among the optimized training sets. In method 2 the thought of early stop in BP neural network was adopted. The optimized training set was acquired firstly by optimizing the training set which set the goal of minimizing the mse to the validation set, secondly by making the prediction to the testing set, and lastly by choosing the one that gave the best prediction results among the optimized training sets. These two methods were used to deal with the simulation data and the gasoline blending data collected from a refinery. The results showed that the methods could improve the accuracy of prediction greatly. These two methods are worth generalizing in chemical engineering modeling in the future.

Key words: particle swarm optimization; optimization of training set; simulation; chemical engineering modeling

2007-07-18 收到初稿, 2007-09-17 收到修改稿。

联系人及第一作者: 张运陶 (1948—), 男, 教授。

基金项目: 四川省企业信息化项目计划 (2005-199-16)。

Received date: 2007-07-18.

Corresponding author: Prof. ZHANG Yuntao. **E-mail:** nczyt

@yahoo.com.cn

引言

对于化工过程,无论是实施生产优化或软测量,其模型的建立均需要大量的前期生产数据来构造训练集。由于化工生产数据受生产波动或仪表精度的影响,一般都不可避免地带有噪声,因此在变量选定后,尽可能剔除原始数据中的“假数据”或无代表性的数据,选出具有代表性的样本构成训练集,是提高优化模型或软测量模型预测准确性的关键。

微粒群^[1] (particle swarm optimization, PSO) 是一种新的进化优化算法,该算法具有易编程、参数调整少、算法收敛速度快等突出优点。本文将 PSO 算法结合机器学习方法用于化工建模,提出两种优化训练集的新思路。一种是应用 PSO,基于交叉验证^[2] (cross validation, CV) 以及对验证集预测,尽可能地剔除原始训练集中的含噪样本,实现训练集优化。该思路实现建模需要一个原始训练集和一个验证集。另一种是借鉴提高 BP 神经网络泛化能力的初期终止^[3] (early stop) 法,在原始训练集和验证集的基础上再增加一个测试集。处理时仍应用 PSO 算法,基于验证集以及测试集双重预测检验,获得能尽可能剔除原始训练集中含噪样本的优化训练集。经过仿真实验和对实际生产数据的研究表明,本文提出的思路可以较大程度提高模型预测准确性,在化工生产优化和软测量建模中具有推广应用价值。

1 PSO 算法优化训练集的设计思路

1.1 微粒群算法简介

以最小优化问题为例,对 PSO 算法^[4-5] 原理简介如下。

设微粒总数为 q , 搜索空间为 n 维。用 $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$ 、 $V_i = (v_{i1}, v_{i2}, \dots, v_{in})$ 、 $P_i = (p_{i1}, p_{i2}, \dots, p_{in})$ 和 $P_g = (p_{g1}, p_{g2}, \dots, p_{gn})$ 分别表示微粒 i ($0 < i \leq q$) 的当前位置、飞行速度、所经历的最好位置和所有微粒经历过的最好位置。其中初始 X 、 V 随机产生, P_i 、 P_g 则通过适应度函数 $f(x)$ 计算确定。有关计算式如下

$$P_i(t+1) = \begin{cases} P_i(t) & \text{if } f(X_i(t+1)) \geq f(P_i(t)) \\ X_i(t+1) & \text{if } f(X_i(t+1)) < f(P_i(t)) \end{cases}$$

$$P_g(t) \in \{P_1(t), P_2(t), \dots, P_i(t), \dots, P_q(t)\} \mid f(P_g(t)) = \min\{f(P_1(t)), f(P_2(t)), \dots, f(P_i(t)), \dots, f(P_q(t))\}$$

微粒群体通过微粒飞行速度和位置的更新来搜

索整个解空间。微粒飞行速度 V 和位置 X 的更新规则为

$$V_i(t+1) = \omega V_i(t) + c_1 r_1(t)(P_i(t) - X_i(t)) + c_2 r_2(t)(P_g(t) - X_i(t))$$

$$V_i(t+1) = \begin{cases} V_i(t+1) & \text{if } V_i(t+1) \leq V_{\max} \\ V_{\max} & \text{if } V_i(t+1) > V_{\max} \end{cases}$$

$$X_i(t+1) = X_i(t) + V_i(t+1)$$

$$X_i(t+1) = \begin{cases} X_i(t+1) & \text{if } X_{\min} < X_i(t+1) < X_{\max} \\ X_{\max} & \text{if } X_i(t+1) \geq X_{\max} \\ X_{\min} & \text{if } X_i(t+1) \leq X_{\min} \end{cases}$$

其中, t 表示迭代次数; ω 为惯性权重, 取定值或呈线性减少; c_1 、 c_2 为加速常数, 通常在 $0 \sim 2$ 之间取值; r_1 、 r_2 为两个相互独立的生成函数, 产生在 $(0, 1)$ 之间的随机数。

算法程序结束条件为找到最优解 (期望输出与寻优结果之差小于某一阈值 ϵ) 或达到最大迭代次数 $t_{\max}$ 。

1.2 PSO 优化训练集的设计思路

本文提出的化工建模思路所用的机器学习方法可根据具体问题选择, 本文采用最小二乘支持向量机^[6] (least squares support vector machine, LSSVM), 以 PSO-LSSVM 算法介绍优化训练集的设计思路, LSSVM 的核函数取 RBF。

1.2.1 设计思路 1

(1) 建立原始训练集和验证集

取装置平稳运行时的前期生产记录, 凭经验剔除异常样本后, 以部分可靠性更高的样本数据作为验证数据集, 其余样本构成原始训练集。

(2) 确定交叉验证折数 k

用 LSSVM 对原始训练集进行交叉验证, 确定最佳交叉验证折数 k 。将 k 折作为模型优化训练集的交叉验证折数。记录对原始训练集 k 折交叉验证的均方误差函数 mse (mean of squared errors) 值。

(3) 寻优运算, 筛选本次模型运行的优化训练集

① 叠代运算 用 PSO 对原始训练集随机抽样建立新的训练集, 以 LSSVM 对新训练集做交叉验证得到的 mse 作为 PSO 算法的适应度函数 fitness 用于指导 PSO 优化训练集, 优化方向为 fitness 值减小的方向。

$$\text{fitness} = \text{mse} = \sum_{i=1}^L (YC - Y_{c_{cv}})^2 / L \quad (1)$$

式中 YC 、 $Y_{c_{cv}}$ 和 L 分别为对应训练集目标值、 k

折交叉验证预测值及样本数。

若 PSO 筛选训练集的 mse 值小于原始训练集的 mse 值, 则作为备选优化训练集。

② 将 mse 最小的备选优化训练集确定为本次模型运行的优化训练集。

③ 对选出的优化训练集再做 k 折交叉验证, 按 mse 最小进一步调节 LSSVM 算法参数, 以尽可能提高模型预测准确性。记录调节得到的算法最佳参数。

(4) 验证比较, 确定建模用的最优训练集。

对 PSO-LSSVM 算法 n 次平行运行获得的 n 个优化训练集, 用 LSSVM 在其最佳参数条件下对验证集预测, 取预测效果最好的训练集作为最优训练集用于化工建模。

对验证集的预测效果评价, 可根据具体的化工问题确定。

1.2.2 设计思路 2

(1) 建立原始训练集、验证集和测试集

类似于思路 1, 从剔除异常样本数据后的前期生产记录中, 选择部分可靠性更高的样本数据分别作为验证集和测试集, 其余的样本构成原始训练集。比较两种思路, 显然思路 2 需要更多的前期生产记录数据。

(2) 以原始训练集用 LSSVM 对验证集做预测, 记录预测结果的均方误差函数 mse 值。

(3) 寻优运算, 筛选本次模型运行的优化训练集

① 叠代运算 用 PSO 对原始训练集随机抽样建立新训练集, 对新训练集以 LSSVM 对验证集做预测, 以预测结果的 mse 作为适应度函数 fitness 值用于指导 PSO 优化训练集, 优化方向为 fitness 值减小的方向。

$$\text{fitness} = \text{mse} = \sum_{i=1}^L (Y_C - Y_c)^2 / h \quad (2)$$

式中 Y_C 、 Y_c 和 h 分别为验证集的目标值、模型对验证集预测值及样本数。

若 PSO 筛选的训练集的 mse 值小于原始训练集的 mse 值, 则作为备选优化训练集。

② 从备选优化训练集中, 选取 mse 最小者为本次模型运行的优化训练集。

③ 对选出的优化训练集再用 LSSVM 对验证集做预测, 进一步按 mse 最小调节 LSSVM 算法参数。记录调节得到的算法最佳参数。

(4) 测试比较, 确定建模用的最优训练集。

对 PSO-LSSVM 算法 n 次平行运行获得的 n 个优化训练集, 用 LSSVM 在其最佳参数条件下对测试集进行预测, 取其中预测效果最好的训练集作为最优训练集用于化工建模。

与思路 1 相同, 对测试集的预测效果评价, 也需要根据具体的化工问题确定。

本思路是借鉴提高 BP 神经网络泛化能力的初期终止 (early stop) 法所作的设计。BP 网络的初期终止法是将前期生产记录数据分为训练集、验证集和测试集 (可选项), 以网络对验证集训练时 mse 值最小为训练目标, 以防止产生“过度训练”, 测试集则用于对训练结果的评价。

2 仿真实验研究

本文采用如下非线性系统作为标准函数^[7]

$$y = \sin(\pi x_1 x_2) + (x_3 - 0.5)^2 + x_4 x_5 \quad (3)$$

对 5 个输入变量 $\{x_i \mid i=1, 2, 3, 4, 5\}$ 取 $[0, 1]$ 之间的随机数并计算对应的 y 值, 分别生成相应的干净数据集。本文中 PSO 算法参数为: 微粒总数 $q=20$, 加速常数 $c_1=c_2=1.8$, 惯性权重 ω 由 1.0 线性减少到 0.4, 选择概率 $r=0.30$, 最大迭代次数 $t_{\max}=100$, 阈值 $\epsilon=0$ 。

2.1 数据集的建立

按上文所述方法, 对于思路 1 分别生成 200 组干净训练集 (clean training set)、100 组验证集 (validation set) 和 100 组用于检验模型的预测集 (prediction set)。通过对干净训练集随机加噪形成 10% 的噪声样本构成含噪原始训练集——噪声训练集 (noise training set)。为了深入讨论, 预测集除干净预测集 (clean prediction set) 外增加一个含噪预测集 (noise prediction set)。含噪预测集是通过对干净预测集随机加噪形成 10% 的噪声样本构成。

对于思路 2, 则在思路 1 的基础上增加 100 组测试集 (testing set)。

2.2 仿真计算

对思路 1, 经研究比较确定对训练集采取 10 折交叉验证。对于两种设计思路, PSO-LSSVM 算法优化训练集均平行运行 10 次。计算时, 先按照设计思路确定最优训练集并建立模型, 再以此模型分别对干净预测集和含噪预测集进行预测。

针对采用的标准函数的具体情况, 本文在两种思路的测试比较步骤, 以及对预测集预测时, 采用 mse 和预测相对误差绝对值 ARE (absolute value

of relative error,%) 分别大于 5 和 10 的个数作为模型效果的双重评价标准,并且将 ARE 作为对预测效果评价的首选标准;即模型预测结果 ARE 分别大于 5%和 10%的个数越少,其预测效果越好。此处,ARE 的计算式为

$$ARE = (|YC - Y_c| / |YC|) \times 100\%$$

(4)

2.3 结果及讨论

2.3.1 干净训练集、噪声训练集和最优训练集的建模情况比较 表 1 给出了按思路 1 执行时干净、噪声和最优 3 种训练集交叉验证及其对验证集的预测结果。对干净、噪声和最优训练集,LSSVM 的模型参数分别为 $\gamma=5450, \sigma=15; \gamma=2700, \sigma=555; \gamma=1050, \sigma=20$ 。按思路 1 执行得到的最优训练集包含 175 个样本,剔除的噪声样本为 13 个、干净样本 12 个,剔除的噪声样本占噪声训练集中噪声样本的 65%。

表 2 给出了按思路 2 执行时 3 种训练集对验证

集和测试集的预测结果。对干净、噪声和最优训练集,LSSVM 的模型参数分别为: $\gamma=7450, \sigma=20; \gamma=4800, \sigma=20; \gamma=6500, \sigma=20$ 。按思路 2 执行时得到的最优训练集包含 166 个样本,剔除的噪声样本为 15 个、干净样本 19 个,剔除的噪声样本占噪声训练集中噪声样本的 75%。

从表 1、表 2 的建模计算结果和以上对最优训练集剔除噪声样本情况的介绍,可以看出:(1) 思路 1 的交叉验证和对验证集的预测,思路 2 对验证集和测试集的预测,都是干净训练集效果最佳,最优训练集其次,而噪声训练集最差。(2) 最优训练集的计算结果明显优于噪声训练集。(3) 两种思路建立的最优训练集,均能有效地剔除噪声样本。但针对仿真实验,思路 2 剔除的干净样本偏多一些。

2.3.2 干净训练集、噪声训练集和最优训练集的预测结果比较 表 3 给出了分别按思路 1 和思路 2 执行时,3 种训练集对干净和含噪预测集的预测结果。

表 1 干净训练集、噪声训练集和最优训练集交叉验证及其对验证集的预测结果

Table 1 Results of cross validation and predicted results to validation set by clean, noise and optimized training set based on method 1

Training set	10-fold cross validation		To validation set		
	mse	R_{cv}^2	mse	$ARE \geq 5\%$	$ARE \geq 10\%$
clean	4.668×10^{-4}	0.997	7.979×10^{-4}	12	2
noise	0.0567	0.666	0.0381	73	61
optimized	0.0080	0.953	0.0044	35	16

表 2 干净训练集、噪声训练集和最优训练集对验证集和测试集的预测结果

Table 2 Predicted results to validation and testing set by clean, noise and optimized training set based on method 2

Training set	To validation set			To testing set		
	mse	$ARE \geq 5\%$	$ARE \geq 10\%$	mse	$ARE \geq 5\%$	$ARE \geq 10\%$
clean	6.339×10^{-4}	11	3	2.555×10^{-4}	9	6
noise	0.0091	62	37	0.0094	55	29
optimized	0.0010	24	7	8.8264×10^{-4}	22	5

表 3 干净训练集、噪声训练集和最优训练集对预测集的预测结果

Table 3 Predicted results to clean and noise prediction set by clean, noise and optimized training set based on method 1 and method 2

Method	Training set	To clean prediction set			To noise prediction set		
		mse	$ARE \geq 5\%$	$ARE \geq 10\%$	mse	$ARE \geq 5\%$	$ARE \geq 10\%$
1	clean	3.49×10^{-4}	8	2	0.0665	16	10
	noise	0.0332	81	59	0.0442	81	59
	optimized	0.0030	20	11	0.0661	26	18
2	clean	2.90×10^{-4}	7	3	0.0695	15	11
	noise	0.0107	52	30	0.1276	56	38
	optimized	5.11×10^{-4}	17	5	0.0693	24	13

分析表 3 数据可见：(1) 对于干净预测集，不论是从预测结果的 mse，还是从 ARE 看，3 种训练集的预测效果均为干净训练集最佳，最优训练集其次，而噪声训练集最差。(2) 思路 2 的预测效果优于思路 1，这与其建模时所用的验证集和测试集均为干净数据集有关。(3) 对于含噪预测集，则出现 mse 和 ARE 反应预测效果不一致的情况。从预测结果 ARE 看，3 种训练集的预测效果仍是干净训练集最佳，最优训练集其次，而噪声训练集最差；从 mse 看，两种思路均有干净训练集和最优训练集接近，而噪声训练集在思路 1 中表现优于干净训练集和最优训练集，在思路 2 中表现则明显较干净训练集和最优训练集差。由于此仿真计算将 ARE 作为预测效果评价的首选标准，因此 3 种训练集对干净和含噪预测集的预测效果均为干净训练集最佳，最优训练集其次，而噪声训练集最差。

3 种训练集对含噪预测集的预测出现 mse 和 ARE 对预测结果反应不同的原因产生于建模方法：思路 1 通过对训练集自身交叉验证训练，再对干净验证集预测建模，因含噪声的训练集建模时会保留部分噪声信息，其结果可导致对含噪预测集预测结果的 mse 偏小；思路 2 训练集通过对干净验证集和干净测试集训练建模，建模过程中含噪声的训练集的噪声信息基本会被剔除干净，因此含噪声的训练集对含噪预测集预测结果的 mse 会偏大。

上述分析表明：(1) 两种思路均能实现对原始噪声训练集的优化；(2) 选择干净数据构成验证集（思路 1）或验证集和测试集（思路 2）用于训练建模，对于提高模型预测效果很重要。因此，在建模时，应注意尽可能使验证集或验证集和测试集数据可靠。

本文 PSO-LSSVM 10 次平行仿真运行的其他 9 次运算获得的优化训练集，对干净和含噪预测集

的预测效果均也优于噪声训练集（限于篇幅不再在此列出结果），表明本文提出的思路完全能够实现优化训练集的目标。

3 实际应用示例

以思路 1 在处理某炼油厂油品车间 93[#] 汽油调和和生产数据中的应用为例作介绍。

3.1 数据集的建立

取剔除异常样本后的前期完整生产记录 198 组，按生产日期取前 140 组作为原始训练集 (original training set)，后 58 组构成验证集；以近期 48 组生产数据构成预测集。按文献 [8] 以组分 1 初馏点、10% 馏点、干点、辛烷值、调和比例，组分 2 初馏点、10% 馏点、干点、辛烷值以及添加剂添加量共 10 个参数构成变量集，调和汽油的辛烷值 (RON) 作为目标值。

根据生产实际，调和汽油生产时模型预测误差绝对值 AE (absolutely error) 应尽可能控制在 0.3 以内，平均绝对误差 MAE (mean of absolutely error) 应尽可能小。因此，此处将 $AE \geq 0.3$ 的样本个数和 MAE 作为预测效果评价标准。

3.2 结果讨论

按思路 1，交叉验证折数确定为 10 折，PSO-LSSVM 算法优化训练集平行运行 20 次，最终选定由 111 组样本构成的最优训练集，运行时 LSSVM 算法最优参数为 $\gamma = 44000$ ， $\sigma = 9020$ 。对于原始训练集运行时 LSSVM 算法最优参数为 $\gamma = 250$ ， $\sigma = 7950$ 。PSO 算法参数同仿真实验。

表 4 给出了原始训练集和最优训练集建模及对预测集的预测结果。

由表 4 可见，无论是建模表现，还是预测结果，训练集优化后得到的最优训练集都明显优于原始训练集。最优训练集与原始训练集比较，其对预

表 4 原始训练集和最优训练集建模及对预测集的预测结果
Table 4 Results of cross validation & predicted results to validation and prediction set by original and optimized training set based on method 1

Training set	10-fold cross validation		To validation set		To prediction set	
	mse	R_{cv}^2	MAE	$AE \geq 0.3$	MAE	$AE \geq 0.3$
original	0.0762	0.0811	0.273	23	0.369	25
optimized	0.0487	0.239	0.220	15	0.237	11

Note: $AE = | \text{measured octane number} - \text{predicted octane number} |$; $MAE = \sum | \text{measured octane number} - \text{predicted octane number} | / \text{number of sample of prediction set}$.

测集的预测结果, MAE 由 0.369 下降到 0.237, $AE \geq 0.3$ 的个数则由 25 下降到 11。应用结果表明, 本文思路能够有效地解决实际应用中的训练集优化问题, 提高模型的预测准确性。

4 结 论

本文提出了两种用 PSO 算法优化化工建模训练集的思路, 并采用 PSO-LSSVM 算法对仿真数据进行了深入的研究及分析讨论。研究表明, 本文提出的两种思路均能有效剔除原始训练集中无代表性的数据, 实现化工建模训练集的优化; 在建模时, 选择干净数据构成验证集 (思路 1) 或验证集和测试集 (思路 2) 用于训练建模, 对于提高模型预测效果十分重要。将思路 1 用于处理某炼油厂汽油调和数据, 应用结果表明, 用对原始训练集优化得到的最优训练集建模, 其预测准确性明显高于用原始训练集建模的预测结果。

References

- [1] Kennedy J, Eberhart R. Particle swarm optimization//Proc. IEEE Int'l. Conf. on Neural Networks (IV). Piscataway, NJ: IEEE Service Center, 1995: 1942-1948
- [2] Chen Nianyi (陈念贻), Qin Pei (钦佩), Chen Ruiliang

(陈瑞亮), Lu Wencong (陆文聪). Application of Pattern Recognition Methods in Chemistry and Chemical Industry (模式识别方法在化学化工中的应用). Beijing: Science Press, 2000: 48-49

- [3] Wen Xin (闻新), Zhou Lu (周露), Li Xiang (李翔), Zhang Baowei (张宝伟). Neural Networks Simulation and Its Application Based on Matlab (神经网络仿真与应用). Beijing: Science Press, 2003: 280-281
- [4] Zeng Jianchao (曾建潮), Jie Qian (介倩), Cui Zhihua (崔志华). Particle Swarm Optimization Algorithm (微粒群算法). Beijing: Science Press, 2004: 12-53
- [5] Shi Y, Eberhart R C. A modified particle swarm optimizer//Proceedings of the IEEE International Conference on Evolutionary Computation. Piscataway, NJ: IEEE Press, 1998: 69-73
- [6] Suykens J A K, De Brabanter J, Lukas L, *et al.* Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, 2002, **48** (1): 85-105
- [7] Yan Weiwu (阎威武), Chang Junlin (常俊林), Shao Huihe (邵惠鹤). Modeling method based on support vector machines within the Bayesian evidence framework. *Control and Decision* (控制与决策), 2004, **19** (5): 525-528
- [8] Zhang Yuntao (张运陶), Duan Qibin (段启彬), Bai Chunyan (白春艳), Yang Xiaoli (杨晓丽), Zheng Wei (郑伟). Computer assisted gasoline blending design. *Journal of Chemical Industry and Engineering (China)* (化工学报), 2006, **57** (7): 1732-1735