DOI:10.15923/j.cnki.cn22-1382/t.2017.5.09

PSO 多元自适应回归样条算法

王盛慧， 赵二卫

（长春工业大学 电气与电子工程学院，吉林 长春 130012）

摘 要： 多元自适应回归样条建模中，样本数据最小步长和端点的设置会影响建模精度。提出了应用粒子群算法优化这两个参数的方法，以预测样本均方差为适应度函数，通过优化最小步长和端点位置调整采样点选取。实例应用表明，PSO-MARS 方法能提高预测精度。

关键词： 粒子群优化；最小步长；端点；交叉验证

中图分类号： TP 391.9 **文献标志码：** A **文章编号：** 1674-1374(2017)05-0459-05

Division method of MARS sample optimized by PSO

WANG Shenghui， ZHAO Erwei

(School of Electrical & Electronic Engineering, Changchun University of Technology, Changchun 130012, China)

Abstract: In the multi-adaptive regression spline modeling process, the setting of both sampling minimum step size and endpointmay influence the precision of modeling. To optimize the two parameters, Particle Swarm Optimization (PSO) method is applied to estimate the Mean Square Error (MSE). The MSE is taken as the fitness function to optimize the minimum step size and endpoint by adjusting the sampling position. Application results indicate that themethod can improve the modeling accuracy.

Key words: Particle Swarm Optimization (PSO); minimum step size; endpoind; cross-validation.

0 引 言

多元自适应回归样条法(Multivariate Adaptive Regression Spline, MARS)是一种专门针对高维数据拟合的回归方法[1-2]。因其建模速度快，可解释性强得到广泛的应用[3]。该方法以样条函数的张量积作为基函数，自动选择插入基函数的节点，构成基函数集合来逼近样本数据。MARS 算法自提出后，很多学者做了研究和改进。由 Friedman 提出的 Fast MARS 算法能在略微降低模型精度的同时加快建模速度。Sergey Bakin[4-5]等提出的 BMARS 使用了并行算法，加快建模速度，同时使模型变得光滑。

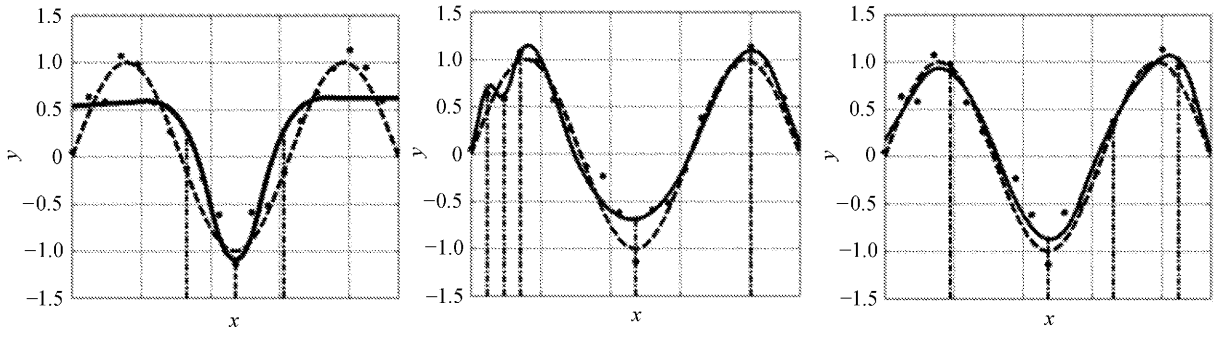
但是当样本数据存在一定干扰时，MARS 可能在干扰点处插入基函数，建立的模型会贴近干扰点，后向剪枝过程不能删除这样的基函数，导致模型在干扰点附近的预测能力下降。

3 种划分方法拟合曲线如图 1 所示。

收稿日期： 2017-06-15

基金项目： 吉林省科技发展计划基金资助项目(20150203003SF)

作者简介： 王盛慧(1976－)，女，汉族，吉林长春人，长春工业大学副教授，硕士，主要从事数字传动与电力节能技术方向研究，E-mail:wangshenghui@ccut.edu.cn.



(a) Minspan=3, Endspan=7　　(b) Minspan=2, Endspan=2　　(c) Minspan=1, Endspan=1

图1　3种划分方法拟合曲线

图中，虚线为正弦曲线，黑色点为正弦函数加上随机扰动，实线为拟合曲线，点画线标识基函数插入点。对同一组数据，采用3种不同的节点划分方法，节点设置参数与模型精度见表1。

表1　3种划分方法拟合参数

图	Minspan	Endspan	实际使用节点数	GCV	MSE
图1(a)	3	7	3	0.226	0.100
图1(b)	2	2	4	0.053	0.017
图1(c)	1	1	5	0.078	0.018

建模过程中，MARS不会处理每个样本点，为了降低局部方差，设置最小步长，用Minspan表示，同时，为了降低数据两侧的局部方差，靠近样本数据两端的点也不会被采用，设置两侧最小放置节点距离，文中用Endspan表示。图1与表1很明显可以看出，Minspan，Endspan过大，采样的数据较少，MARS的拟合能力较差；而图1(c)虽然使用了最多的节点和基函数，Minspan=1，Endspan=1，每一个样本数据都被采样，但是由于插入基函数的节点扰动较大，所以并不能很好地挖掘出正弦关系，从而模型此样本点处的预测能力下降；图1(b)的拟合程度说明，按照Minspan=2，Endspan=2的样本划分方法，MARS能较好地反映出系统的特征。可以看出，样本的划分方法能在很大程度影响模型的精度和预测能力。针对MARS的这种问题，文中提出应用粒子群算法(PSO)来优化MARS样本空间划分方法。

1　MARS算法简介

多元自适应回归样条(MARS)是由Friedman引入的一种回归分析形式，它是一种非参数回归技术，可以看作模拟变量之间的非线性和相互作用的线性模型的扩展。MARS模型的一般形式：

$$\hat{y}=a_0+\sum_{m=1}^{M}a_m\prod_{k=1}^{km}[s_{km}(x_{v(k,m)}-t_{km})] \quad (1)$$

式中：$\hat{y}$——模型输出变量值；

a_0——常数；

a_m——基函数的系数；

$\prod_{k=1}^{km}[s_{km}(x_{v(k,m)}-t_{km})$——基函数；

M——基函数的个数；

$x_v(k,m)$——独立变量的标识；

t_{km}——变量空间的节点。

它的建模过程分为前向逐步和后向剪枝。前向过程是一个迭代过程，模型首先生成初始基函数(迭代次数$I=0$)，即

$$B_0(x)=1 \quad (3)$$

每次迭代($I>1$)，MARS遍历所有节点，成对地添加新的来减小训练误差最多的镜像基函数，直到基函数个数达到最大个数或者模型精度满足要求：

$$B_{2I-1}(x)=B_l(x)b(x_v,t) \quad (4)$$

$$B_{2I}(x)=B_l(x)b(-x_v,t) \quad (5)$$

式中：$B_l(x)$——在之前的迭代中生成的基函数，

称作父基函数。

这种迭代过程会产生大量的基函数，造成模型的过拟合，后向剪枝过程每次循环删除一个对训练误差减小量为最小的基函数，得到对应子模型，直到模型只剩下截距项，引入广义交叉验证 GCV 准则：

$$\mathrm{GCV}=\frac{1}{n}\frac{\sum_{m=1}^{n}(y_i-\hat{y}_i)^2}{\left(1-\frac{C(M)}{n}\right)^2} \tag{6}$$

$$C(M)=\mathrm{trace}(B(B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}})+1+dM \tag{7}$$

式中：M ——基函数个数；

n ——输入变量个数；

$\mathrm{trace}(B\ (B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}})+1$ ——模型有效系数个数；

d ——惩罚因子，一般设为 2 到 4 之间。

最终选取 GCV 值最小的子模型作为最优模型，可以看出过多的基函数与扭结点会受到惩罚，从而减小模型的体积，避免过拟合。模型不会处理所有样本点，引入最小步长 $L(a)$，即 Minspan：

$$\mathrm{Minspan}=L(a)=-\frac{\log_2\left[-\frac{1}{pN}\ln(1-a)\right]}{2.5}$$
$$0.01<a<0.05 \tag{8}$$

式中：P ——样本组数；

N ——输入变量个数。

节点的选取会直接影响模型的精度和复杂度，尤其对有干扰的样本，在干扰点处添加基函数，可能会导致过拟合和预测能力下降，怎样划分样本空间直接影响模型的精度和复杂度。

2　PSO-MARS 算法

2.1　PSO 算法

粒子群算法(PSO)是通过模拟鸟群觅食在解空间中通过迭代搜索出最优解的方法[6-8]，算法首先随机生成粒子群的位置和速度：

$$X_i=(x_{i1},x_{i2},\cdots,x_{iD})^{\mathrm{T}} \tag{9}$$

$$V_i=(V_{i1},V_{i2},\cdots,V_{iD})^{\mathrm{T}} \tag{10}$$

根据粒子的适应度至获取粒子最优位置和全局最优位置

$$P_i=(V_{i1},V_{i2},\cdots,V_{iD})^{\mathrm{T}} \tag{11}$$

$$P_g=(P_{g1},P_{g2},\cdots,P_{gD})^{\mathrm{T}} \tag{12}$$

在下一次迭代中粒子更新自己的位置和速度

$$v_{id}^{k+1}=v_{id}^{k}+c_1r_{1d}(p_{id}^{k}-x_{id}^{k})+c_2r_{2d}(p_{gd}^{k}-x_{id}^{k}) \tag{13}$$

$$x_{id}^{k+1}=x_{id}^{k}+v_{id}^{k+1} \tag{14}$$

式中：d ——维度；

k ——迭代次数；

c_1,c_2 ——学习因子；

r_{1d},r_{2d} ——0 到 1 之间的随机值。

从上式可以看出，粒子具有自我总结和向优秀个体学习的能力，较之于遗传算法，粒子群有记忆能力、操作简单、收敛迅速的特点。

2.2　PSO-MARS 算法

为了能够增加模型的鲁棒性，准确反映系统特征，对样本数据采用 10 折交叉验证的建模方法。确保所有数据都有机会参与模型的训练和预测，算法流程如图 2 所示。

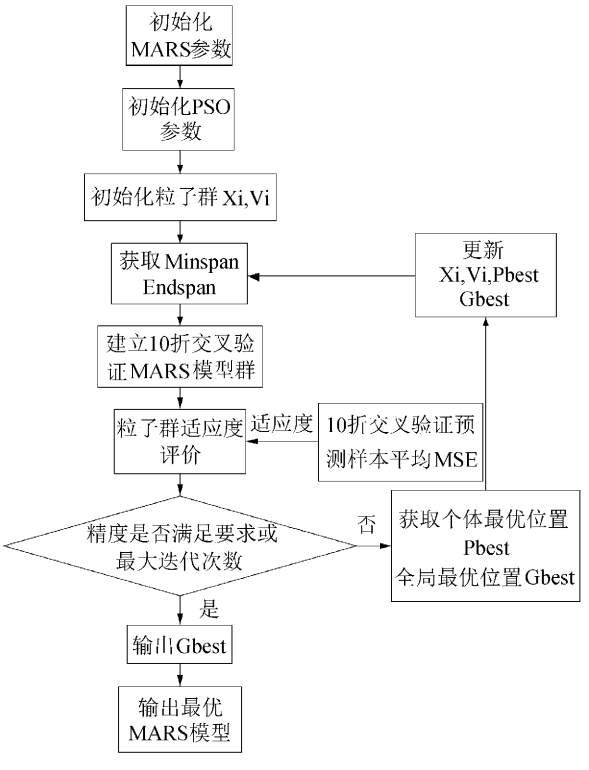


图 2　算法流程

算法步骤如下：

1)初始化 MARS 参数，初始化 PSO 参数。

2)随机生成粒子群 $X_{i\,\mathrm{Minspan}}$，$V_{i\,\mathrm{Minspan}}$，$X_{i\,\mathrm{Endspan}}$，$V_{i\,\mathrm{Endspan}}$。

3)将随机步骤 2)生成的 Minspan 和 Endspan 传递给 MARS，划分样本空间，建立 MARS 模型群。

4)以 10 折交叉验证的预测样本 MSE 平均

值为适应度函数，计算粒子群的适应度值，平均MSE值最低的粒子 X_i 作为全局最优位置，传给 G_{best}，单个粒子在迭代过程中得到最低MSE值的 X_i 作为个体最优位置传给 P_{best}。

5)根据式(13)和(14)，更新粒子群 X_i，V_i，P_{best}，G_{best}，将 X_i 传递给MARS重新划分样本空间，开始新一轮的计算。

如果迭代次数达到设置的最大迭代次数或者模型精度满足要求，即输出最优模型和 G_{best}。

3 实例应用

测试数据来自UCI机器学习数据库，该数据集来自联合循环电厂，以温度、环境压力、相对湿度、排气真空度来预测每小时净电能输出。选取数据集中的800个样本作为训练样本，80个样本作为测试样本。初始化MARS参数，最大基函数个数设置50，最大交互程度设置2；初始化PSO参数，随机生成粒子群

$$X_{i\,\mathrm{Minspan}}=(X_{i1\,\mathrm{Minspan}},X_{i2\,\mathrm{Minspan}},\cdots,X_{in\,\mathrm{Minspan}})$$

$$V_{i\,\mathrm{Minspan}}=(X_{i1\,\mathrm{Minspan}},X_{i2\,\mathrm{Minspan}},\cdots,X_{in\,\mathrm{Minspan}})$$

$$X_{i\,\mathrm{Endspan}}=(X_{i1\,\mathrm{Endspan}},X_{i2\,\mathrm{Endspan}},\cdots,X_{in\,\mathrm{Endspan}})$$

$$V_{i\,\mathrm{Endspan}}=(X_{i1\,\mathrm{Endspan}},X_{i2\,\mathrm{Endspan}},\cdots,X_{in\,\mathrm{Endspan}})$$

其中，n 为种群数，n 取10，迭代次数设置为40次，Minspan和Endspan为两个不同的粒子群，迭代时并行计算，为了加快搜索速度，初始最小步长在式(8)范围内随机生成。

PSO优化Minspan和Endspan过程如图3所示。

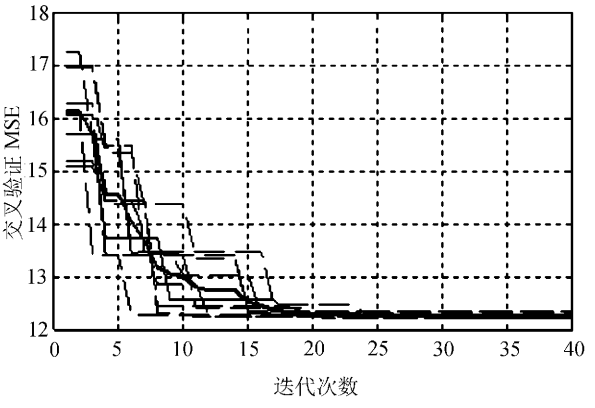


图3 优化过程

从图3可以看出，随着迭代次数的增加，交叉验证MSE和平均MSE都在下降，验证间的初始MSE不同，但都有一定的下降空间。

同时，做了另外两组仿真来对比，参数设置及仿真结果见表2。

表2 PSO-MARS结果对比

Minspan	Endspan	训练		预测	
		GCV	MSE	MSE	ME
$L(0.25)$	10	16.88	14.73	15.70	16.81
1	1	17.09	14.43	15.46	17.33
PSO	PSO	17.33	14.88	12.27	12.68

3组仿真的训练精度差别不大，Minspan=1，Endspan=1时，模型遍历每个节点，但训练和预测精度并不是最高。使用PSO搜索的划分方法Minspan为9，Endspan为2，训练精度与其它两组基本一致，但预测精度和最大偏差有显著提高，相对其他两组，平均MSE分别降低了21.8%和20.6%。

选取其中一折预测拟合图形，如图4所示。

从图4可以看出，3组预测效果都很好，PSO优化的一组相对整体更加贴近样本数据。

4 结 语

以带扰动的正弦函数为例，设置3组不同的最小步长和端点，模型精度和预测能力差别很大，说明这两个参数对多元自适应回归样条算法有很大影响。

针对手动设置最小步长和端点往往不能取得最优值的问题，文中提出用PSO来优化这两个参数的方法，优化的适应度函数为预测精度，同时采用交叉验证的建模方法来增加模型的鲁棒性，给出了详细的结合算法步骤。

将 PSO-MARS 应用与联合循环电厂电能输出建模中，从 MSE 收敛曲线可以看出，经过 PSO 的优化，预测 MSE 有一定幅度的减小，PSO-MARS 可以在训练精度基本不变的情况下，提高预测精度，可以用于对建模速度要求不高的离线模型建立中，有一定的实际意义。

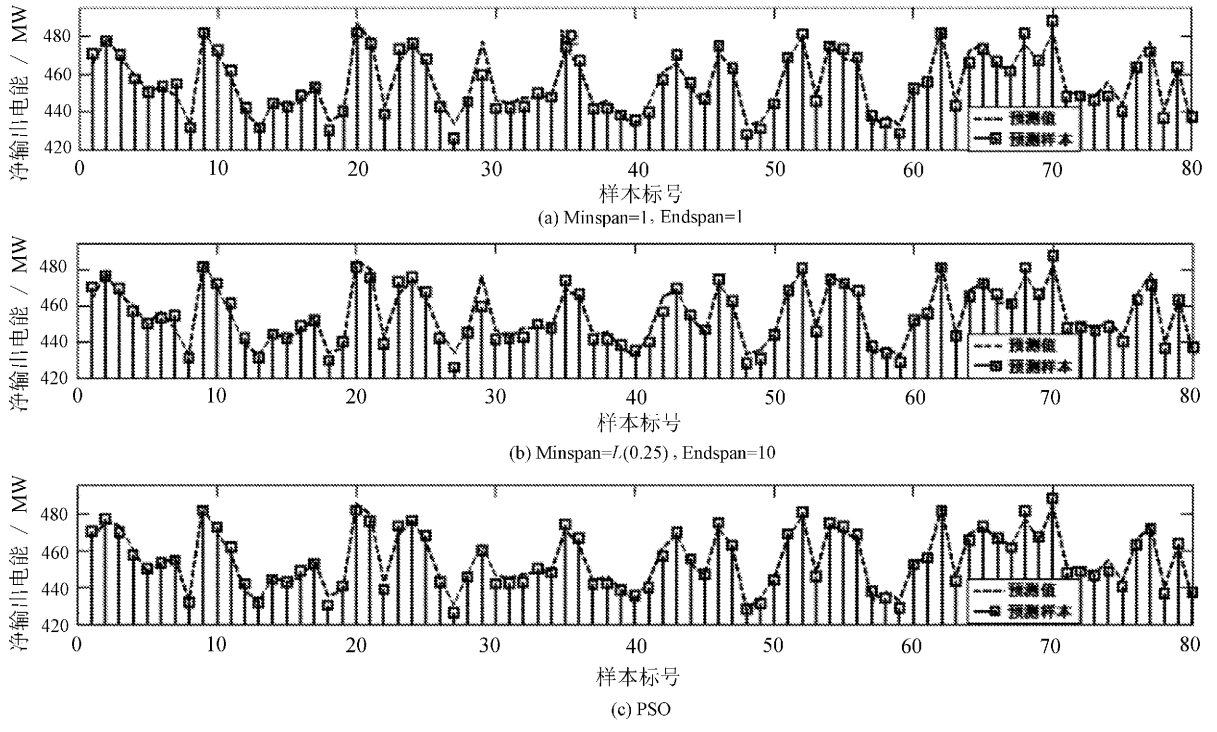


图 4　3 种预测拟合图形

参考文献：

[1] Friedman J H. Multivariate adaptive regression splines (with discussion) [J]. The Annals of Statistics, 1991, 19(1): 123-141.

[2] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction[M]. 2nd edition. [S.l.]: Springer, 2009.

[3] 宋阳，凌震华，戴礼荣.基于合成质量预测的单元挑选语音合成优化方法[J].清华大学学报：自然科学版，2013(6)：762-766.

[4] Bakin S, Hegland M, Osborne M. Can MARS be improved with B-splines? [M]. New Jersey, USA: Computational Techniques and Applications Conference, 1998: 75-82.

[5] 初众，吴义忠，陈立平，等.基于黄金分割法的加速 MARS 研究[J].系统仿真学报，2012(8)：1561-1566.

[6] 冯非凡，武雪玲，牛瑞卿，等.粒子群优化 BP 神经网络的滑坡敏感性评价[J].测绘科学，2017(10)：1-9.

[7] 邱东，刘明硕，郭红涛.基于粒子群算法的低碳铬铁磷含量预测研究[J].计算机技术与发展，2017(6)：1-4.

[8] 金星，徐婷，冷淼.基于 IPSO-SVR 的水泥分解炉温度预测模型研究[J].现代电子技术，2017(9)：148-151.