

DOI: 10.7511/jslx20201110001

基于 k -fold 交叉验证的代理模型序列采样方法

李正良^{1,2}, 彭思思¹, 王涛^{*1}

(1. 重庆大学 土木工程学院, 重庆 400045;

2. 重庆大学 山地城镇建设与新技术教育部重点实验室, 重庆 400045)

摘要:在代理模型序列采样框架下,针对现有研究中的不足之处,通过引入 k -fold 交叉验证计算样本的预测误差,并结合泰森多边形法和最大距离最小化准则,发展了一种适用于任意代理模型的 k -fold CV-Voronoi 自适应序列采样方法。相较于传统序列采样方法,本文方法具有计算简单和自适应性强等显著优势。通过数值算例和工程算例对比分析发现所提序列采样方法具有较高的近似精度和计算效率,此外,进一步讨论了 k -fold 交叉验证中 k 的不同取值对于代理模型精度的影响,总结出 k 的最优取值范围以供参考。

关键词: k -fold 交叉验证; 序列采样; 代理模型; 泰森多边形

中图分类号: O213 **文献标志码:** A **文章编号:** 1007-4708(2022)02-0244-06

1 引言

对于一些较为复杂的实际工程问题,常采用物理实验和仿真模拟进行深入探究和分析,但这些方法往往需要消耗大量的人力物力,耗时长,成本高,其应用受到很大限制。而代理模型技术可以有效解决上述问题,成为一种非常流行的代替昂贵数值模拟的有效方式^[1]。代理模型的本质是用输入和输出参数来模拟响应系统的行为,通过选取一定数量的样本点来构建具有相当精度的数学函数模型,其在工程设计和优化中得到广泛应用^[1]。作为构建代理模型的关键步骤,采样方法至关重要,采样方法选取的样本点很大程度上决定了代理模型的精度水平。如何选取尽可能少的样本点构建具有足够精度的代理模型一直是研究者们关注的重点和热点,对此国内外学者展开了大量的研究^[1]。

现有的代理模型采样方法一般分为一次性采样方法和序列采样方法两大类。一次性采样(one-shot sampling)方法主要考虑空间填充设计,使样本点尽量均匀地分布在整个设计空间中,较为常见的方法有全因子试验设计法、正交设计法和拉丁超立方采样法等^[2]。然而在实际工程中采样数量往往难以确定,一次性采样容易导致过采样或欠采样^[2],造成计算资源的浪费或近似精度难以满足

要求,且自适应能力差。相较于一次性采样方法,序列采样方法(sequential sampling)对初始样本的依赖性较弱,能有效控制样本的数量,可以根据实际物理系统的空间特性进行采样,因此得到了学者们的青睐^[3-7]。序列采样方法的基本思路为选取少量样本初步构建近似模型,在序列迭代的过程中充分利用已有样本和模型的信息来确定候选样本点的位置,通过在欠采样和高度非线性的局部区域逐步加点,不断提高模型的近似精度,直至满足设计者的需求。当前,该方法已应用于模型预测^[3,4]、优化设计^[5,6]以及可靠度分析^[7]等诸多领域。

为了在整个设计空间中得到良好的全局近似结果,兼顾代理模型的全局精度和局部精度,不少学者针对全局代理模型的序列采样方法进行了深入探究。其中,泰森多边形(或 Voronoi)作为一种合理划分空间和分析影响区域的常用工具,能有效地对整个设计空间进行全局探索,近些年来应用于序列采样方法的相关研究中。Crombecq 等^[2]提出了一种 LOLA-Voronoi 序列采样方法,创新性地根据泰森多边形法衡量样本密度,并通过局部线性逼近法分析数据点周围的非线性程度,进而选择新的候选样本点。但该方法在估计样本的梯度信息时需进行复杂的邻域选择,计算速度较慢;Xu 等^[8]结合了泰森多边形法和 LOO 交叉验证(leave-one-out cross validation)误差进行采样,发展了一类鲁棒性强的 CV-Voronoi 序列采样方法;缪继华等^[9]将 CV-Voronoi 方法与支持向量回归模型相结合,提出了基于支持向量点的空间加点策略。虽然

收稿日期:2020-11-10;修改稿收到日期:2021-03-09.

基金项目:国家自然科学基金(51478064);国家自然科学基金国际(地区)合作与交流(51611140123)资助项目.

作者简介:王涛*(1993-),男,博士生
(E-mail:taowang@cqu.edu.cn).

CV-Voronoi 序列采样方法具有较高的精度和较强的鲁棒性,但其序列采样过程中交叉验证涉及大量的建模计算,且随着样本点数量的增加,计算量不断上升,致使其计算成本较为昂贵。

鉴于此,为了节约计算成本、加快采样速度及提高建模效率,本文引入 k -fold 交叉验证 (k -fold cross validation)^[10] 方法,并结合泰森多边形法对设计空间进行全局探索和局部开发,发展了一种 k -fold CV-Voronoi 自适应序列采样方法,并通过数值算例和工程算例对本文方法进行验证。

2 序列采样方法

2.1 k -fold CV-Voronoi 序列采样方法

本文发展的 k -fold CV-Voronoi 自适应序列采样方法基于泰森多边形法剖分设计空间,通过 k -fold 交叉验证方法展开误差追踪,在交叉验证误差最大的区域内进行自适应采样。该方法核心内容包括剖分设计空间、确定子区域以及选取候选点。

(1) 设计空间剖分。泰森多边形由离散点连成的三角形各条边的垂直平分线围成,是对空间区域的一种无缝不重叠剖分。对于 d 维设计空间 Ω 的一组样本点 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, 其中第 i 个点 \mathbf{x}_i 的 Voronoi 区域 D_i 定义为^[9]

$$D_i = \{ \mathbf{x} \in \Omega \mid d(\mathbf{x}, \mathbf{x}_i) < d(\mathbf{x}, \mathbf{x}_j) \} \quad (j = (1, 2, \dots, n), j \neq i) \quad (1)$$

式中 n 为已知的样本数量, \mathbf{x} 为设计空间 Ω 的向量点, $d(\mathbf{x}_i, \mathbf{x}_j)$ 为样本点 \mathbf{x}_i 和 \mathbf{x}_j 之间的距离。

以二维设计空间为例,在 $[-1, 1]^2$ 范围内随机生成 10 个样本点,其 Voronoi 单元划分情况如图 1 所示。可以看出,任意一个泰森多边形 D_i 内部只存在一个已知样本点 \mathbf{x}_i , 且多边形 D_i 中的点与其中心样本 \mathbf{x}_i 的距离最近,因此可将多边形 D_i 围成的区域视为样本 \mathbf{x}_i 的有效影响范围。

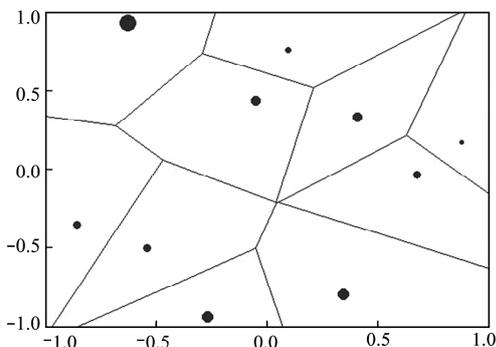


图 1 二维 Voronoi 图及其单元预测误差

Fig. 1 2D Voronoi diagram and the prediction errors

(2) 子区域确定。当整个设计空间划分为一系列多边形区域后,需进一步考虑代理模型在各个子区域内的近似情况,在预测误差较大的区域增加样本点能有效提高模型精度。交叉验证方法可以获取样本点的预测误差信息,可用于评价代理模型的近似效果,但常见的 LOO 交叉验证方法随着序列采样过程的进行,所需的计算量在成倍上升,计算成本较大。因此,本文引入 k -fold 交叉验证方法^[10] 计算样本点的预测误差,其详细步骤如图 2 所示。

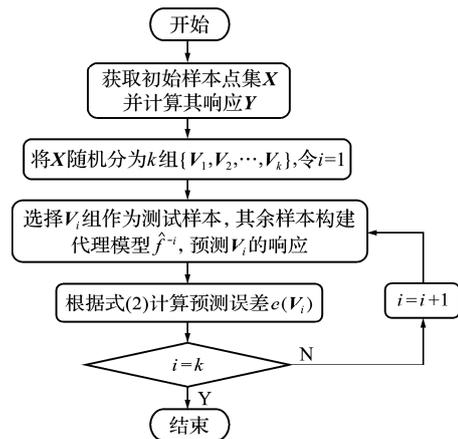


图 2 k -fold 交叉验证方法流程

Fig. 2 Flowchart of k -fold cross validation method

k -fold 交叉验证方法将样本集 \mathbf{X} 随机分成 k 组 ($k \leq n$) 互斥的子集 $\{\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_k\}$, 分别计算各集合中样本点的预测误差^[1], 其表达式可写为

$$e(\mathbf{V}_i) = |f(\mathbf{V}_i) - \tilde{f}^{-i}(\mathbf{V}_i)| \quad (i = 1, 2, \dots, k) \quad (2)$$

式中 $f(\mathbf{V}_i)$ 为第 i 个集合 \mathbf{V}_i 样本的实际响应值, \tilde{f}^{-i} 为由除集合 \mathbf{V}_i 以外其他已知样本点构建的代理模型, $\tilde{f}^{-i}(\mathbf{V}_i)$ 为 \mathbf{V}_i 所有样本的预测响应值。

图 1 所示的二维实例中,采用不同大小的实心圆点表示由式(2)计算的样本点预测误差值,当样本点 \mathbf{x}_i 的误差 e_i 越大,即表明在多边形 D_i 区域内代理模型的拟合效果越差,与实际函数存在较大偏差。假设第 c 个样本点 \mathbf{x}_c 的误差最大,在相应的多边形 D_c 区域内进行采样。

(3) 候选点选取。为保证样本的空间填充特性,最大程度获取函数局部信息, k -fold CV-Voronoi 自适应序列采样法应满足多边形区域 D_c 内的候选样本点与其中心点 \mathbf{x}_c 之间的距离尽可能大。具体操作为在预测误差最大的多边形区域内产生一批随机点,根据式(3)所示的最大距离最小化 (maxmin) 准则^[11] 抽取新的样本点,

$$\max [d(\mathbf{x}, \mathbf{x}_c)] \quad (\mathbf{x} \in D_c) \quad (3)$$

式中 $\max(\cdot)$ 为求最大值的函数, \mathbf{x} 为区域 D_c 的向量点。

对于二维空间,采用上述方法进行采样的结果如图3所示,其中空圆圈代表一系列随机点,实心点即为新增样本点。

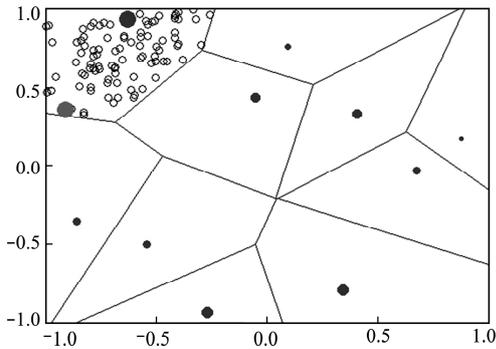


图3 二维采样过程
Fig. 3 2D sampling process

与一般的序列采样方法对比,本文建议的 k -fold CV-Voronoi 自适应采样方法存在三点优势。其一,该采样方法本身完全独立于代理模型技术,因此能与多种类型的代理模型相结合,可应用于不同工程问题,具有普适性;其二,算法中包括探索阶段和开发阶段两个过程,通过泰森多边形的空间划分和交叉验证误差的计算,实现了整个设计空间内局部开发和全局探索之间的动态平衡,自适应性强;其三,引入 k -fold 交叉验证策略能有效减少建模过程的计算成本,加快采样速度,提高建模效率。

2.2 算法流程

对于实际工程中的建模问题,采用 k -fold CV-Voronoi 自适应序列采样方法构建精确代理模型的具体步骤如下。

(1) 选取适当的试验设计方法,在设计空间中抽取少量初始样本集 \mathbf{X} , 并计算其响应 \mathbf{Y} 。

(2) 根据样本及其响应值,建立相应的代理模型,并计算模型的近似精度。

(3) 基于已有样本信息,利用泰森多边形法剖分整个设计空间。

(4) 采用 k -fold 交叉验证方法计算每个样本点的预测误差,选出误差最大的样本点所在的多边形区域。

(5) 在选出的多边形区域中产生一系列随机点,根据 maxmin 准则选择候选样本点 \mathbf{x}_{new} 并计算真实响应值 y_{new} , 将其加入集合 \mathbf{X} 和 \mathbf{Y} , 更新样本。

(6) 检查停止准则。若计算结果满足给定的停止准则,建模结束;否则,继续步骤(2),进行下一

轮迭代。

3 数学算例

RBF 是目前应用最为广泛的代理模型之一^[12], 其选择某一欧氏距离函数作为基函数,通过径向基函数的线性组合来拟合实际函数,所得模型近似精度比较高,具有良好的鲁棒性,适用于解决高度非线性问题^[13]。为验证 k -fold CV-Voronoi 自适应序列采样方法的效率以及适用性,将建议的自适应序列采样方法与 RBF 代理模型相结合,应用于数学函数算例进行对比分析,并针对 k -fold 交叉验证方法中 k 的取值展开讨论。

3.1 数学函数测试

将建议的 k -fold CV-Voronoi 自适应序列采样方法应用于 5 个典型的数学函数,建立 RBF 模型并检验其精度,并与 Xu 等^[8]提出的基于 LOO 交叉验证的 CV-Voronoi 采样方法进行对比。数学测试函数的相关信息列入表 1。

表 1 测试函数信息
Tab. 1 Information of test function

| Function number | Function name | Range of values | dimension |
|-----------------|----------------------------|-----------------|-----------|
| F_1 | Zakharov ^[15] | $[-5, 5]$ | 2 |
| F_2 | Rastrigin ^[15] | $[-1, 1]$ | 2 |
| F_3 | Hartman ^[15] | $[0, 1]$ | 3 |
| F_4 | Rosenbrock ^[15] | $[-5, 5]$ | 3 |
| F_5 | Sphere ^[15] | $[-5, 5]$ | 6 |

为了获得尽可能多的函数信息,初始样本点应在整个设计空间中均匀分布,并且具有良好的投影特性,故本文采用操作简单和易于实现的拉丁超立方抽样法(LHS)^[14]选取初始样本。在序列加点过程中,考虑有限的计算成本,设置停止准则为,循环加点过程中当生成的总样本达到设置的最大样本数量时停止采样。若测试函数的设计维度为 d , 设置初始样本数量为 $N_0 = 5d$, 最大样本个数为 $N = 4N_0$ 。抽取 500 个测试样本验证代理模型的精度,采用两个常见的模型全局精度指标评价模型的近似程度,均方根误差(RMSE)和平方相关系数(R^2)的表达式为

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m |f(\mathbf{x}_i) - \tilde{f}(\mathbf{x}_i)|} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^m [f(\mathbf{x}_i) - \tilde{f}(\mathbf{x}_i)]^2}{\sum_{i=1}^m [f(\mathbf{x}_i) - \bar{f}]^2} \quad (5)$$

式中 m 为测试样本点数量, $f(x_i)$ 为测试样本点的实际响应值, $\bar{f}(x_i)$ 为实际响应的平均值, $\tilde{f}(x_i)$ 为代理模型在测试样本点处的预测值。

为了消除随机误差的影响,表 2 所列结果为多次计算的平均值。

表 2 不同采样方法下 RBF 建模结果
Tab.2 Modeling results of different sampling methods

| Function number | Cross validation method | RMSE | R^2 | Num |
|-----------------|-------------------------|----------|----------|------|
| F_1 | k -fold($k=5$) | 34.97310 | 0.954002 | 155 |
| | k -fold($k=5d$) | 38.47190 | 0.947440 | 310 |
| | LOO | 53.91639 | 0.900845 | 735 |
| F_2 | k -fold($k=5$) | 0.257363 | 0.913829 | 155 |
| | k -fold($k=5d$) | 0.215287 | 0.946969 | 310 |
| | LOO | 0.242081 | 0.938519 | 735 |
| F_3 | k -fold($k=5$) | 0.354774 | 0.861423 | 230 |
| | k -fold($k=5d$) | 0.325987 | 0.881339 | 690 |
| | LOO | 0.403389 | 0.813805 | 1665 |
| F_4 | k -fold($k=5$) | 8279.683 | 0.893499 | 230 |
| | k -fold($k=5d$) | 9141.206 | 0.868599 | 690 |
| | LOO | 8796.631 | 0.881728 | 1665 |
| F_5 | k -fold($k=5$) | 4.835629 | 0.046455 | 455 |
| | k -fold($k=5d$) | 4.699334 | 0.045130 | 2730 |
| | LOO | 4.629350 | 0.045185 | 6705 |

注: Num 为交叉验证计算过程中产生的建模次数。

由表 2 可知,在初始样本和总样本数量不变的情况下,本文方法所得能以少量的计算成本取得较好的建模效果,而 LOO 交叉验证策略下序列采样过程中的建模次数远高于 k -fold 交叉验证。设初始样本数量为 N_0 , 采样结束时的样本总量为 N , 采用 LOO 交叉误差验证的 CV-Voronoi 自适应序列加点方法共需要 $1/2(N+N_0-1)(N-N_0)$ 次建模,而采用本文的 k -fold CV-Voronoi 方法仅需 $k(N-N_0+1)$ 次建模计算,大幅缩减了计算消耗。

另一方面,对比表 2 各个函数模型的精度指标发现,在计算成本较小的前提下,本文方法的精度水平不低于 CV-Voronoi 自适应序列加点方法。综合而言,本文方法计算成本低,计算效率高,且鲁棒性强。

3.2 关于 k 的讨论

值得注意的是,表 2 中代理模型的精度结果随本文方法 k 的取值变化而发生改变,因此,为了确定 k 值的取值,本文进一步探究不同 k 值对代理模型精度的影响。仍采用表 1 的函数为例,对于每个测试函数,分别选取了不同的初始样本值 N_0 , 其相应 k 值在 N_0 的公因数按从小到大的顺序分别进行取值^[16],如当初始样本个数 $N_0=12$ 时, k 值分别取 2,3,4,6 和 12,样本总量保持不变。采用本文方法建立 RBF 模型,以全局指标 R^2 衡量模型精度,将多次测试结果整理成如图 4 所示的箱线图。

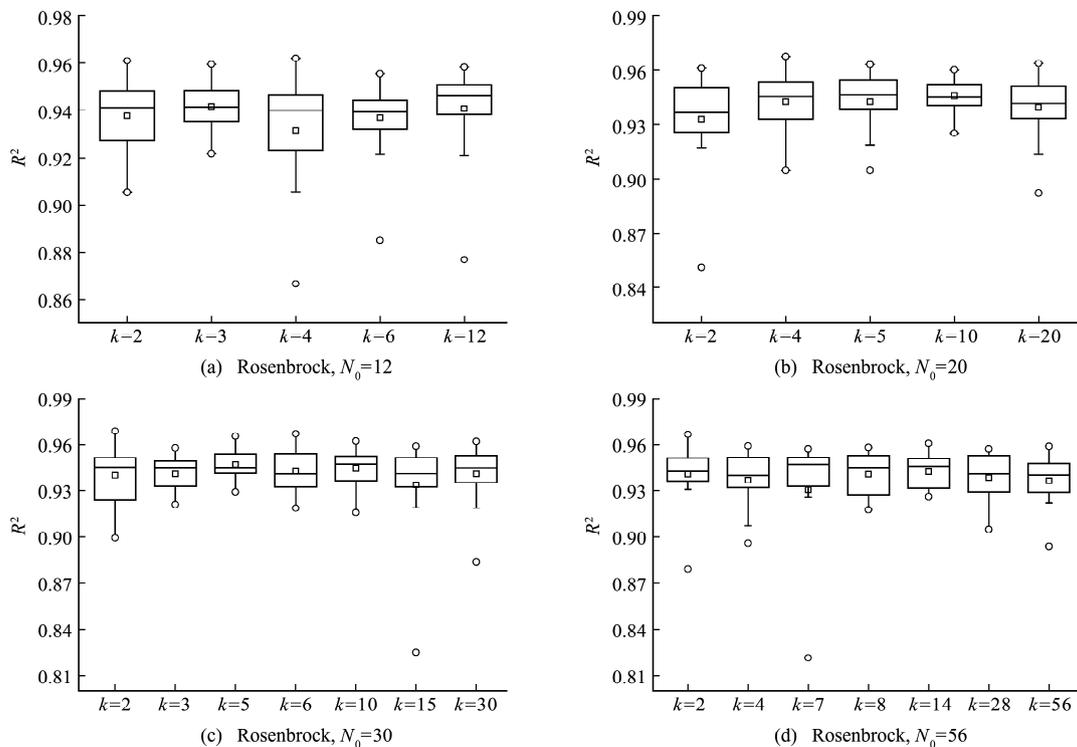


图 4 不同 k 值情况下 RBF 模型精度
Fig.4 Accuracy of RBF model with different k values

通过观察图4平均数、众数和分位数等数据发现,在初始样本数量相同的情况下,同一测试函数不同 k 值下所得代理模型计算结果均未表现出明显的线性规律,即当 k 发生变化时,RBF模型的整体精度 R^2 可能上升亦可能下降,在局部区域会出现极值。根据试验中多种情况下所得测试结果可知,对于所有测试函数,当 $k/N_0=1/6\sim 1/3$ 时,所得模型的精度较高,且一般在 $k=1/4 * N_0$ 处取得极值,建模效果最佳(如图4中Rosenbrock函数在 $N_0=12, k=3; N_0=56, k=14$ 时取得最优结果),并且此时交叉验证的次数较少,计算成本低。

文献[17,18]也对比了LOO和 k -fold交叉验证,证明在一次性采样建模中 k -fold交叉验证方法取 $k=10$ 时取得较优结果。但采用序列采样方法构建代理模型时,可以通过不断添加新的样本来更新模型,在多次循环迭代过程中不断提高模型精度,常选用较小的初始样本值和 k 值,故在上述 k -fold CV-Voronoi方法测试中, k 的建议取值略小于以往文献给出的参考值。

4 工程算例

图5所示的交叉斜材作为保证输电塔结构稳定性的重要部件,对承受横向荷载起着关键作用,并且常与辅材一起改善输电塔内部的结构受力。

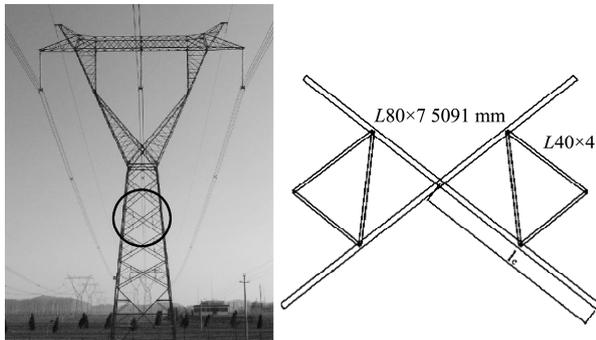


图5 输电塔交叉斜材构造

Fig. 5 Structure of diagonal cross-bracing in transmission tower

交叉斜材的承载力与截面几何尺寸和拉压比等因素之间的函数关系是其设计与优化的关键。本文选取交叉斜材的角钢截面肢长 l 、肢宽 b 以及拉力与压力的比 γ 作为自变量,通过ANSYS有限元分析软件对输电塔交叉斜材进行数值模拟分析,计算不同参数下杆件的最大内应力响应,采用本文方法建立RBF模型。建模过程中初始样本数量为 $N_0=15$,测试样本数量为30, k -fold交叉验证的 k 值取为 $1/4 N_0$ 。

表3展现了建模过程中随着样本数量 N 的增加,RBF模型精度的变化情况。由表3可知,本文

方法能快速提高代理模型的近似程度,当样本总量为30时,代理模型的整体精度 R^2 达到80%以上,能够以较小的计算成本取得令人满意的建模效果,且所得代理模型可进一步用于输电塔交叉斜材的设计和优化。在实际应用中,设计者可根据目标函数精度要求控制加点过程的进行,从而提高建模效率,节约了计算成本。

表3 RBF模型精度
Tab. 3 Accuracy of RBF model

| | $N=15$ | $N=20$ | $N=30$ | $N=75$ |
|-------|---------|---------|---------|---------|
| RMSE | 25.9246 | 23.6548 | 19.8411 | 16.4605 |
| R^2 | 0.70455 | 0.78987 | 0.82971 | 0.87810 |

5 结论

针对现有采样方法的不足之处,本文发展了一种适用于任意代理模型的 k -fold CV-Voronoi自适应序列采样方法,该法能有效地在整个函数设计空间进行全局探索和局部开发,具有很强的自适应性和鲁棒性。通过数学算例和工程算例得出结论如下。

(1) 相同条件下,本文方法在序列迭代过程中交叉验证的建模计算次数较低,能大幅减少计算成本,提高建模效率。

(2) 本文方法适用于复杂系统的建模分析,且在计算资源有限的情况下具有显著优势。

(3) 提供了本文方法中 k 值的最佳取值范围,即 $k=(1/6\sim 1/3)N_0$,可供设计及科研人员参考。

参考文献(References):

- [1] Liu H T, Ong Y S, Cai J F. A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design[J]. *Structural and Multidisciplinary Optimization*, 2018, **57**(1): 393-416.
- [2] Crombecq K, Gorissen D, Deschrijver D, et al. A novel hybrid sequential design strategy for global surrogate modeling of computer experiments[J]. *SIAM Journal on Scientific Computing*, 2011, **33**(4): 1948-1974.
- [3] Mackman T J, Allen C B. Investigation of an adaptive sampling method for data interpolation using radial basis functions[J]. *International Journal for Numerical Methods in Engineering*, 2010, **83**(7): 915-938.
- [4] Lehmensiek R. Efficient Adaptive Sampling Applied to Multivariate, Multiple Output Rational Interpolation Models, with Applications in Electromagnetics-based Device Modelling[D]. University of Stellenbosch, 2001.
- [5] Regis R, Shoemaker C. A stochastic radial basis func-

- tion method for the global optimization of expensive functions[J]. *Informs Journal on Computing*, 2007, **19**:497-509.
- [6] 刘克龙,姚卫星,穆雪峰.基于 Kriging 代理模型的结构形状优化方法研究[J]. *计算力学学报*, 2006, **23**(3): 344-347, 362. (LIU Ke-long, YAO Wei-xing, MU Xue-feng. A method of structural shape optimization based on Kriging model[J]. *Chinese Journal of Computational Mechanics*, 2006, **23**(3): 344-347, 362. (in Chinese))
- [7] Xiang Z L, Chen J H, Bao Y Q, et al. An active learning method combining deep neural network and weighted sampling for structural reliability analysis [J]. *Mechanical Systems and Signal Processing*, 2020, **140**:106684.
- [8] Xu S L, Liu H T, Xiao F W, et al. A robust error-pursuing sequential sampling approach for global meta-modeling based on Voronoi diagram and cross validation[J]. *Journal of Mechanical Design*, 2014, **136**: 071009.
- [9] 缪继华.基于支持向量回归模型的序列采样方法研究[D]. 华中科技大学, 2017. (MIAO Ji-hua. A Sequential Sampling Method Based on Support Vector Regression[D]. Huazhong University of Science and Technology, 2017. (in Chinese))
- [10] Meckesheimer M, Booker A J, Barton R R, et al. Computationally inexpensive metamodel assessment strategies[J]. *AIAA Journal*, 2002, **40**(10): 2053-2060.
- [11] Johnson M E, Moore L M, Ylvisaker D. Minimax and maximin distance designs[J]. *Journal of Statistical Planning and Inference*, 1990, **26**(2):131-148.
- [12] Fang H B, Horstemeyer M F. Global response approximation with radial basis functions[J]. *Engineering Optimization*, 2006, **38**(4):407-424.
- [13] Fang H, Rais-Rohani M, Liu Z, et al. A comparative study of metamodeling methods for multiobjective crashworthiness optimization[J]. *Computers & Structures*, 2005, **83**(25-26):2121-2136.
- [14] McKay M D, Beckman R, Conover W. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code [J]. *Technometrics*, 2000, **42**:55-61.
- [15] <http://www.sfu.ca/~ssurjano/optimization.html>.
- [16] Viana F A C, Haftka R T, Steffen V. Multiple surrogates: how cross-validation errors can help us to obtain the best predictor[J]. *Structural and Multidisciplinary Optimization*, 2009, **39**(4):439-457.
- [17] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[J]. *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, **14**:1137-1145.
- [18] Brieman L, Spector P. Submodel selection and evaluation in regression: The x-random case[J]. *International Statistical Review*, 1992, **60**(3):291-319.

A sequential sampling method of surrogate model based on k -fold cross validation

LI Zheng-liang^{1,2}, PENG Si-si¹, WANG Tao^{*1}

(1. School of Civil Engineering, Chongqing University, Chongqing 400045, China;

2. Key Laboratory of New Technology for Construction of Cities in Mountain Area, Ministry of Education, Chongqing University, Chongqing 400045, China)

Abstract: Under the framework of sequence sampling of surrogate models, in view of the shortcomings of existing methods, a k -fold CV-Voronoi adaptive sequential sampling method was developed, which is suitable for arbitrary surrogate models. In this method, the k -fold cross-validation was introduced to calculate the prediction error of sample points, and the Voronoi diagram and maxmin criterion were combined. Compared with the traditional sequential sampling method, the proposed method has the advantages of calculation simplicity and strong adaptability. Through the numerical examples and engineering example, it is shown that the proposed sequential sampling method has high accuracy and calculation efficiency. In addition, the influence of different k values on the accuracy of surrogate model is further discussed in the k -fold cross-validation, and the optimal range of k values is summarized for reference.

Key words: k -fold cross validation; sequential sample; surrogate model; Voronoi diagram

引用本文/Cite this paper:

李正良,彭思思,王涛.基于 k -fold 交叉验证的代理模型序列采样方法[J]. *计算力学学报*, 2022, **39**(2):244-249.

LI Zheng-liang, PENG Si-si, WANG Tao. A sequential sampling method of surrogate model based on k -fold cross validation[J]. *Chinese Journal of Computational Mechanics*, 2022, **39**(2):244-249.