

基于增量 SVR 模型的改进全局优化算法

贾云峰¹ 张爱莲² 吴义忠¹

(华中科技大学国家 CAD 支撑软件工程技术研究中心 武汉 430074)¹

(武汉纺织大学机电学院 武汉 430073)²

摘 要 SVR(支持向量回归)是一种具有较强稳健性的小样本学习方法,它可有效避免“维数灾难”,并被引入到全局优化中。然而现有的基于 SVR 的全局优化算法存在估值次数多、无法应对高维优化问题等缺点。提出了一种基于增量 SVR 模型的新的改进全局优化算法 DISVR:采用增量 SVR 方法提高过程响应面的重构效率;采用一种新的增量 LHD(Latin Hyper-cube Sampling)方法确保样本集分布均匀;采用 DIRECT 搜索算法提高全局搜索的稳定性和效率。最后,通过多个测试函数表明,该算法既降低了时间复杂度,也有效减少了源模型的估值次数。

关键词 全局优化,响应面,支持向量回归,增量法

中图分类号 TP391 **文献标识码** A

Improved Global Optimization Algorithm Based on Incremental Support Vector Regression Model

JIA Yun-feng¹ ZHANG Ai-lian² WU Yi-zhong¹

(National Enterprise Information Supported Software Engineering Centre, Mechanical Department of

Huazhong University of Science and Technology, Wuhan 430074, China)¹

(College of Mechanical and Electrical Engineering, Wuhan Textile University, Wuhan 430073, China)²

Abstract SVR(Support Vector Regression) is a kind of small sample learning method with strong robustness. It can effectively avoid ‘dimension disaster’, and is introduced to the global optimization. However, the existing global optimization algorithms based on SVR have several shortcomings, such as large number of evaluations, can not cope with high dimensional optimization problem and so on. We proposed a new improved global optimization algorithm DISVR based on incremental SVR model; an incremental SVR method to improve the efficiency of reconstruction process response, a new incremental LHD sampling(Latin Hyper-cube Sampling) to ensure an uniform distribution of samples, DIRECT search algorithm to enhance the stability and efficiency of the global search. Finally, the result of test functions suggests that the proposed method both reduce the time complexity, also effectively reduce number of source model’s evaluations.

Keywords Global optimization, Response surface, Support vector regression, Incremental method

1 引言

目前存在很多传统全局优化算法,有通过生成一个确定序列逼近收敛全局最优点的确定性方法,如分支定界方法(branch and bound)、聚类方法(clustering)和隧道方法(tunneling method)等;有基于一系列随机生成的可行点进行局部优化的遗传算法(GA)、模拟退火方法(SA)、统计算法(statistic algorithm)、禁忌搜索方法(Tabu Search)、蚁群算法(ACO)和粒子群算法(PSO)等^[1,2]。近来出现了一种基于元模型(响应面)的全局仿真优化算法,该方法是以试验设计与数理统计为基础的函数逼近类全局优化方法,通过较少的试验来获得设计变量和设计目标之间足够准确的相互关系,利用响应面技术显著减少计算成本。常用作响应面的有 RSM、Kriging、RBF 和 SVR 等元模型(meta-model)^[2]。传统全局优化方法及其它基于替代模型的全局仿真优化方法存在估值次

数多、无法应对高维优化问题等缺点^[3]。

支持向量机(Support Vector Machines, SVM)由 Vapnik 领导的 AT&T Bell 实验室研究小组于 1963 年提出,是一种基于统计学习理论的分类技术^[4]。支持向量回归(SVR, Support Vector Regression)基于 SVM 理论,构造源模型的响应面。现有的基于 SVR 响应面的全局优化方法无法保证样本数很小时选择出具有代表性的样本,使之覆盖整个设计区间;重构 SVR 模型时间较长,最优解搜索速度较慢;不能有效应对约束条件下的全局寻优^[5]。

本文提出了一种基于增量 SVR 模型的全局优化算法 DISVR。算法采用南安普顿大学 S. R. Gunn 教授用 MATLAB 程序编写的 SVM 工具包^[6],选择 RBF 径向基函数作为建模核函数,采用 DIRECT(Divided Rectangle)法作为响应面寻优搜索方式^[7]。所提算法集中于采用一种新的增量 LHD 采样方法,以确保样本集分布均匀;建立近似结构分析模型

到稿日期:2011-05-17 返修日期:2011-09-16 本文受国家自然科学基金(50775084, 50975107)资助。

贾云峰(1988—),男,硕士生,主要研究方向为多领域仿真优化, E-mail: jyf565732259@163.com; 张爱莲(1973—),女,讲师,主要研究方向为 CAD 和优化设计等; 吴义忠(1970—),男,教授,主要研究方向为多领域仿真优化等。

时,为快速优化大批量采样点,利用支持向量机理论自身特点,根据支持向量集与训练样本集之间存在的等价关系,构建一种增量 SVR 算法重构响应面。因支持向量具有良好的泛化能力,解决了算法复杂度与输入向量维数密切相关的问题,故理论上可对任何大型结构进行近似建模;采用加约束的 DIRECT 算法作为搜索策略求解结构模型,可有效解决带约束优化问题。

2 SVR 模型及其算法流程

2.1 SVR 模型

支持向量回归也分为线性回归和非线性回归两种,但不是统计学中的线性或者非线性回归,而是根据是否需要嵌入到高维空间来划分的,简述如下^[4]。

对于给定的样本集 S 以及任意给定的 $\epsilon > 0$,如果在原始空间 R^n 存在超平面

$$y = w \cdot x + b, w \in R^n, b \in R \quad (1)$$

使得

$$|y_i - f(x_i)| \leq \epsilon, \forall (x_i, y_i) \in S \quad (2)$$

则称 $y = w \cdot x + b$ 是样本集合 S 的 ϵ -线性回归。

式(2)等价于 S 中任何点 (x_i, y_i) 到超平面式(1)的距离不超过 $\frac{\epsilon}{\sqrt{1 + \|w\|^2}}$ 。调整超平面的斜率 w , 使其与 S 中任

一点 (x_i, y_i) 距离都尽可能大,即使得 $\frac{\epsilon}{\sqrt{1 + \|w\|^2}}$ 最大化,这等价于要求 $\min \{\|w\|^2\}$ 。于是, ϵ -线性回归问题就转化为优化问题

$$\begin{cases} \min \frac{1}{2} \|w\|^2 \\ \text{s. t. } |\langle w, x_i \rangle + b - y_i| \leq \epsilon, i = 1, 2, \dots, l \end{cases} \quad (3)$$

于是,引入松弛变量,并使用 Lagrange 乘子法,得到优化问题的对偶形式并求解。

假如训练样本在原空间中线性不可分,我们引入非线性变换将原始样本集映射到高维的 Hilbert 空间中。在此空间中进行线性回归,并采用合适的核函数来简化运算,这就是支持向量非线性回归。

此时,回归方程变为

$$y = K(w \cdot x) + b \quad (4)$$

直接输入原始维度的训练样本进行拟合即可。

SVR 常用的核函数多为以下几种。

多项式核

$$K(x_i, x_j) = ((x_i \cdot x_j) + c)^d, \text{其中}, c \geq 0, d \text{ 为任意整数} \quad (5)$$

Gauss 径向基核

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (6)$$

Sigmoid 核

$$K(x_i, x_j) = \tanh(K(x_i, x_j) + v), K > 0, v < 0 \quad (7)$$

2.2 算法流程

Step1 采样:结合随机 LHD 和最小距离最大化采样,构造一种新的增量 LHD 采样方法,选取一个有限的初始点集合 $S_1 = \{x_1, \dots, x_k\} \subseteq D, D$ 为可行域,用于复杂函数估值;

Step2 第一次全局寻优:在初始点集 S_1 上构造 SVR 响应面,将构造的响应面作为带约束 Direct 搜索方法的目标函数进行寻优;

Step3 迭代寻优:将第一次寻优得到的最优点加入到采样集合中,并新增一个到所有采样点最小距离最大的点以及一个随机采样点。在此采样集上构造新的增量 SVR 响应面,进行迭代寻优,当不满足精度要求时,增加采样集合,进行循环迭代。直到满足精度要求或达到循环次数时,退出迭代。

基于增量 SVR 的全局优化算法 DISVR 的流程图如图 1 所示。

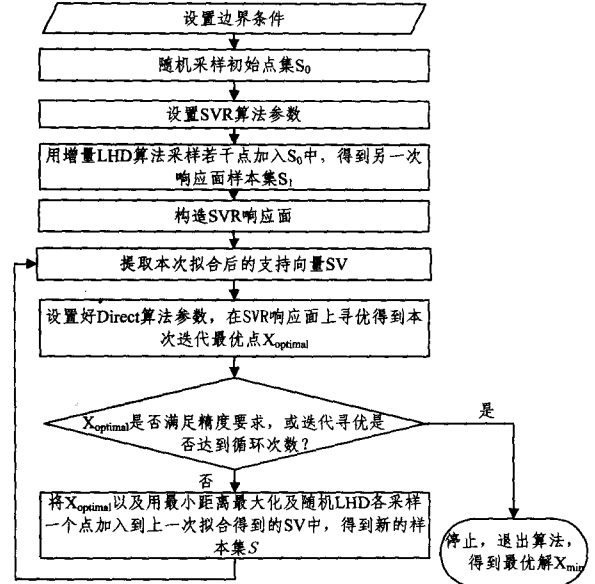


图1 基于增量 SVR 的全局优化方法流程图

3 算法关键过程分析

3.1 增量 LHD 采样方法

此方法建立在随机拉丁超立方采样基础上。随机拉丁超立方采样 (LHD, Latin Hyper-cube Sampling) 属于全空间填充且非重叠的随机采样方法。它的采样点在全局域内是均匀的^[8]。LHD 的主要思想在于,首先将 n 个维度区间进行 m 等分,然后在每个维度内随机取值,再根据每个维度值随机地选取配对,已经选取的不再重复,形成 m 个采样点的 n 维数据。本文提出的增量 LHD 采样方法 (MinmaxLHD) 是使得生成的采样点满足任意两点之间的最小距离最大化。假设现有用 LHD 采得的初始点集为 X_P , 新采点集为 X_C , 总的采点集为 X_A , 即 $X_A = X_P \cup X_C$, 最小距离最大方法要求满足

$$\max_{X_C} \left[\min_{\substack{1 \leq i \leq m, 1 \leq j \leq l+m}} (d(x_G, x_{A_j})) \right] \quad (8)$$

算法采样时,首先取若干点,然后循环一定次数,每次得到的新采样点都是到之前采样点最小距离最大的点。后面全局优化加入新采样点时,同样采用这种方式,得到的样本点均匀,构造出来的响应面模型不会过于集中,有效地避免了拉丁方采样的随机性。

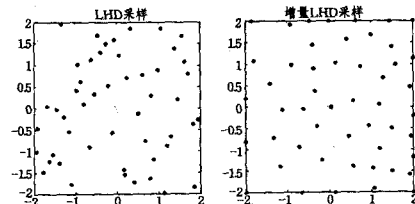


图2 随机 LHD 采样与增量 LHD 采样比较

3.2 全局搜索 DIRECT

DIRECT 是一个自适应细分全局优化算法,由 Donald R.

Jones 提出^[7]。它是对标准 Lipschitz 方法的改进^[9],无需给定 Lipschitz 收敛性常数。Lipschitz 常数是决定对全局或局部搜索的一个平衡参数。DIRECT 算法从所有可能的值中确定该常数,以实现在全局和局部范围内的自适应搜索最优解。DIRECT 算法思路是将设计域看成一个 n 维的超矩形(超立方体),将其分割成 3 等分,再寻找其中目标值最小的矩形继续细分,直到迭代结束^[2]。

基于增量 SVR 模型的全局优化算法中的 DIRECT 法,其目标函数用增量 SVR 构造的响应面代替。

3.3 增量 SVR 方法

SVR 算法稳定,在小样本高维函数中应用优势明显,具有坚实的统计学依据。但由于它的最终决策函数只由少量支持向量所决定,计算的复杂性取决于支持向量的数目,而不是样本空间的维数,因此在构建响应面时可应用增量 SVR 算法。

增量 SVR 算法中,每次的新训练样本由上一次训练得到的支持向量 SV 样本组成,再次训练则只需进行一次即可完成。这样虽忽略了历史样本集的非 SV 最终可能成为后续训练支持向量的问题,因为它减少了计算复杂度,在用替代模型需要多次训练重新构造替代模型时节省了大量的建模时间,对整个优化算法的影响比较小,因而选择此种增量模式。

以下结果为首先采用随机 LHD 采得 50 个点,再循环 3 次采得 150 点,总共 200 点,每次循环过程中构建一次响应面所耗总时间。表 1 为用原始 SVR 与增量 SVR 的 8 个测试函数结果对比。

表 1 原始 SVR 与增量 SVR 测试结果比较

测试函数	SVR		增量 SVR	
	RMSE	时间(s)	RMSE	时间(s)
Alpine Function(AF)	0.1744	101.31161	0.2033	43.442908
Banana Function(BF)	12.9202	86.616673	15.5440	60.302626
Branin function(BR)	0.2055	93.200147	0.2264	40.430477
Schaffer's Function	0.0604	113.559061	0.0622	35.538622
Generalized polynomial function(GF)	2.2259	89.212267	2.4793	53.690038
Six-hump Camel-Back function	0.2327	92.237768	0.2824	41.598936
Rastrigin Function(RS)	1.3127	100.536216	1.4497	36.39068
Himmelblau function(HMD)	59.8043	88.589448	74.1218	54.096938

4 测试结果对比

以下结果都是在各算法的 MATLAB 程序语言中进行的测试。

表 2 中数据为基于增量 SVR 响应面的全局优化算法 DISVR 与传统或其它常用全局仿真优化算法的测试结果比较,使用的是 8 个经典测试函数。其主要着眼于源模型估值次数以及测试所得最优值及最小值与理论结果之间的比较。其中,本文所提的 DISVR 算法的测试结果采取 10 次测试,并取其中间值的方式。

根据测试结果可作如下分析:

- 1) 基于 SVR 响应面的全局优化方法有效地降低了大部分测试函数的源模型估值次数,尤其在多峰函数中优势明显;
- 2) 测试结果更加接近理论最优值,且最小值更为精确,稳定性更高;
- 3) 对一些峰值不明显的“平底锅”型函数的测试中,结果

不显著,有一定偏差。

表 2 DISVR 与传统全局优化算法的测试对比

算法	源函数估值次数	得到的最优值	目标最小值
Alpine Function(AF),理论最优值和最小值为(7.9171,4.8158);-6.1295			
GA	1000	(0.0000,0.0000)	0.0000
SA	1391	(7.9171,4.8158)	-6.1295
PSO	125	(7.9180,4.8046)	-6.1291
DIRECT	267	(7.9218,4.8171)	-6.1294
SKO	220	(7.9171,4.7872)	-6.1269
DISVR	37	(7.9172,4.8171)	-6.1295
Banana Function(BF),理论最优值和最小值为(1,1);0			
GA	1000	(1.0796,1.1656)	0.0063
SA	2153	(0.9531,0.9085)	0.0022
PSO	271	(1.0005,1.0111)	0.0011
DIRECT	1347	(0.9931,0.9858)	0.0007
SKO	120	(0.9884,0.9770)	1.3706e-004
DISVR	61	(1.0151,1.0297)	2.7460e-004
Branin function(BR),最小值为 0.3979,3 个理论最优值为 (- π ,12.275),(- π ,2.275)和(-3 π ,2.475)			
GA	1000	(-3.1418,12.2767)	0.3979
SA	3201	(-3.1416,12.2750)	0.3979
PSO	184	(-3.1403,12.2790)	0.3979
DIRECT	159	(-3.1241,12.2165)	0.3996
SKO	220	(-3.1415,12.2750)	0.3979
DISVR	56	(-3.1401,12.2668)	0.3979
Schaffer's Function,理论最优值和最小值为 (0.0000,0.0000);0.0000			
GA	1000	(0.0041,0.0072)	0.0001
SA	1015	(0.0000,0.0000)	0.0000
PSO	417	(0.0000,0.0000)	0.0000
DIRECT	69	(0.0000,0.0000)	0.0000
SKO	200	(0.0001109,0.0000221)	1.2799e-008
DISVR	2	(-4.9676e-04,4.0644e-04)	4.117e-007
Generalized polynomial function(GF),理论最优值和最小值为 (2.0000,0.1703);0.5233			
GA	1000	(2.0000,0.1701)	0.5233
SA	3465	(2.0000,0.1701)	0.5233
PSO	85	(2.0000,0.1703)	0.5233
DIRECT	7931	(2.0000,0.1701)	0.5234
SKO	220	(2.0000,0.1705)	0.5233
DISVR	78	(2.0000,0.1691)	0.5234
Six-hump Camel-Back function,最小值为-1.0316,两个理论最优值为 (0.08983,0.7216),(-0.08983,0.7216)			
GA	1000	(0.08983,0.7216)	-1.0316
SA	2465	(0.08983,0.7216)	-1.0316
PSO	432	(0.08983,0.7216)	-1.0316
DIRECT	97	(-0.1481,0.7407)	-1.0135
SKO	41	(-0.0869,0.7258)	-1.0301
DISVR	76	(-0.0896,0.7133)	-1.0316
Rastrigin Function(RS),理论最优值和最小值为(0.0000,0.0000);-2			
GA	1000	(0.0011,-0.0008)	-1.9998
SA	1007	(0.0000,0.0000)	-2.0000
PSO	437	(0.0000,-0.0000)	-2.0000
DIRECT	失败		
SKO	75	(0.000725,0.000244)	-1.9999
DISVR	32	(-3.0483e-04,1.0161e-04)	-2.0000
Himmelblau function(HMD),理论最优值和最小值为(3.0000,2.0000);0.0000			
GA	1000	(3.0065,1.9982)	0.0014
SA	1270	(2.9998,2.0001)	9.6808e-007
PSO	366	(3.0002,1.9998)	0.0000
DIRECT	209	(2.9991,1.9832)	5.3088e-005
SKO	220	(2.9960,2.0071)	8.9042e-004
DISVR	124	(3.0002,1.9999)	7.2534e-007

结束语 本文提出了一种基于增量支持向量回归响应面及 DIRECT 搜索的全局优化算法。测试函数的结果表明,本算法在保证建模速度的情况下可有效降低源模型估值次数,

因而在工程实际中具有良好的应用前景。本算法不足之处在于,对一些“平底锅”型函数,无法保证其稳定性。其原因一方面在于增量 SVR 响应面拟合时尚无法做到十分精确,另一方面在于样本采样不能自适应建模情况。

未来的任务是在模型的样本集合中使用结合本算法特点的采样方式,构建一种自适应的响应面搜索方法。

参考文献

- [1] Younis A, Gu J C, Dong Z M, et al. Trends, features, and tests of common and recently introduced global optimization methods [C]// The 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Victoria, Canada, 2008; 691-718
- [2] 吴义忠, 陈立平. 多领域物理系统的仿真优化方法[M]. 北京: 科学出版社, 2011; 248-249
- [3] An Jin-long, Yang Qing-xin, Ma Zhen-ping, et al. A global optimization algorithm based on Support Vector Machines for electromagnetic inverse problem[C]// Automation Congress, 2008. WAC 2008. World, 2008; 1-5

- [4] Chapelle O, Haffner P, Vapnik V N. Support Vector Machines for Histogram-based Image Classification[J]. IEEE Trans. on Neural Networks, 1999, 10(5): 1055-1064
- [5] 吕勇波. 基于支持向量机与遗传算法的结构优化研究[J]. 华中科技大学, 2007
- [6] Gunn S R. Support vector machines for classification and regression[M]. Faculty of Engineering and Applied Science Department of Electronics and Computer Science, 1998; 63
- [7] Jones D R. DIRECT Global optimization algorithm[J]. Encyclopedia of Optimization. Dordrecht; Kluwer Academic Publications, 2001; 25
- [8] Myers R H, Montgomery D C. Response Surface Methodology-Process and Product Optimization Using Designed Experiments [J]. New York: Wiley Press, 2002; 308
- [9] Jones D R, Perttunen C D, Stuckman B E. Lipschitz optimization without Lipschitz Constant[J]. Journal of Optimization Theory and Applications, 1993, 79(1): 201-210

(上接第 167 页)

部分的时间和结点数与边数存在正相关关系,但是平均度数不是重要影响因素。这与我们对 k -s 算法扰乱部分的时间复杂度的分析是一致的。

可以注意到,扰乱部分的运行时间要比将图 G 划分成 k 个离散的连通子图的时间快一个或者多个数量级。也就是说,整个算法的运行时间主要取决于后者。这提醒我们,如果能高效地将图 G 划分成 k 个离散的连通子图,将极大地改善整个算法的性能。

显然地,总运行时间减去扰乱部分的运行时间,就是将原图划分成 k 个离散的子图的运行时间。我们没有单独给出分析,是因为这部分选用的是参考文献[8]的算法,虽然实验的数据集不同,但结论是一致的。

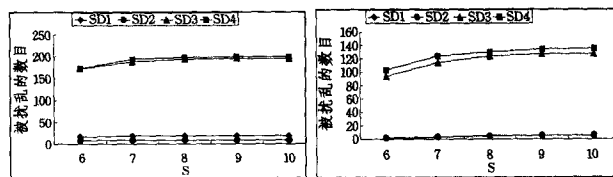


图 10 $k=12$ 时,不同 s 取值时 图 11 $k=20$ 时,不同 s 取值时
扰乱数目(点数和边数) 扰乱数目(点数和边数)

另外,从图 10 和图 11 可以看出, s 改变,被扰乱的点和边的数量没有明显改变。 k 变大,被扰乱的点和边的数量有减少趋势。这是因为 k 越大,点分组和边分组中敏感属性的取值个数可能会增加。可以看见相同 k 和 s 取值的情况下,SD1 和 SD2、SD3 和 SD4 被扰乱的点和边的数量分别比较接近。这里没有发现平均度数的影响。

结束语 本文提出了防止结点泄露、边泄露和特征泄露的三级隐私保护的概念。首次正式提出了社会网络中特征泄露的概念,并对此进行了理论探讨。在此基础上,提出了 k -s 算法作为解决方案,该算法框架具有一定的灵活性。最后,我们在真实数据集 Speed Dating 上进行了实验,实验结果证明

了本算法有效。在后面的工作中,可以将其他找 k 个离散的连通同构子图的高性能算法引入,也可将点或边的敏感属性个数从 1 个扩展到多个。

参考文献

- [1] 兰丽辉,鞠时光,金华. 社会网络数据发布中的隐私保护研究进展[J]. 小型微型计算机系统, 2010, 31(12): 2318-2323
- [2] Hay M, Miklau G, Jensen D, et al. Resisting structural re-identification in anonymized social networks[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 102-114
- [3] Backstrom L, Dwork C, Kleinberg J M. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography[C]// Proceedings of the Sixteenth International World Wide Web Conference, WWW, 2007. Piscataway: IEEE Computer Society, 2007; 181-190
- [4] Zhou Bin, Pei Jian, Luk W-S. A brief survey on anonymization techniques for privacy preserving publishing of social network data[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008. New York: ACM Press, 2008; 12-22
- [5] Liu Kun, Terzi E. Towards identity anonymization on graphs[C]// Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008. New York: ACM Press, 2008; 93-106
- [6] Zhou Bin, Pei Jian. Preserving privacy in social networks against neighborhood attacks[C]// Proceedings of the 24th International Conference on Data Engineering, 2008. Piscataway: IEEE Computer Society, 2008; 506-515
- [7] Zou Lei, Chen Lei, Özsu M T. K-automorphism: A general framework for privacy preserving network publication[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 946-957
- [8] Cheng J, Fu A W-C, Liu Jia. K-isomorphism: privacy preserving network publication against structural attacks[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data, 2010. New York: ACM Press, 2010; 459-470