

自步学习在确定含水层参数中的应用

杨陈东¹, 常安定², 张 明¹

(1. 西安航空学院理学院, 陕西 西安 710077; 2. 长安大学理学院, 陕西 西安 710061)

摘要:通过分析受到误差影响的抽水试验数据进行含水层估计,为含水层参数估计提供方法支持。以机器学习领域的自步学习方法为基础,构造了基于差分进化算法优化的自步学习方法,并将其应用到含水层参数确定中;在不同误差水平下,与其他方法进行对比试验。结果表明,在不同误差水平下,估计参数值与传统估计值之间以及仿真数据与原始数据之间均保持较小差异;基于抽水试验数据估计含水层参数的自步学习方法估计结果有效可靠,算法对误差的稳定容错性强。

关键词:水资源;含水层参数;抽水试验;自步学习;差分进化算法

中图分类号:TV211.1⁺2 **文献标志码:**A **文章编号:**1006-7647(2017)05-0074-04

Application of self-paced learning to estimation of aquifer parameters//YANG Chendong¹, CHANG Anding², ZHANG Ming¹(1. School of Science, Xi'an Aeronautical University, Xi'an 710077, China; 2. School of Science, Chang'an University, Xi'an 710061, China)

Abstract: Through analysis of the pumping test data, which were affected by errors, an aquifer was investigated in order to present a new method for estimating the aquifer parameters. Based on the self-paced learning method in the machine learning field, a self-paced learning method based on the differential evolution algorithm was established and applied to the determination of aquifer parameters. This method was compared with other methods under errors of different levels. The numerical experimental results show that under errors of different levels, the differences between the values estimated with this method and traditional methods are small, and there are minor differences between the simulation data and original data. The results of the self-paced learning method, which was used for estimation of aquifer parameters based on pumping test data, are effective and reliable. The established algorithm is highly stable and it is seldom affected by data errors.

Key words: water resources; aquifer parameters; pumping test; self-paced learning; differential evolution algorithm

在地下水资源评价与开发过程中,导水系数和储水系数等含水层参数是非常重要的基本数据,分析非稳定流抽水试验数据是确定含水层参数的重要途径之一。针对 Sushil 方法及直线图解法等^[1-2]传统方法的局限性,将智能优化算法应用在确定含水层参数中,表现出一定的优越性^[3,4],由于此类方法目标函数均采用最小二乘形式,这里将其统称为最小二乘方法。最小二乘方法均没有考虑实测数据的误差,事实上,误差是普遍存在的,数据在获取和传输过程中都存在误差污染的可能。当数据存在较大误差时,以最小二乘为目标函数的优化问题将严重偏离真实结果。因此,如何从含有比较严重的误差和奇异值的抽水试验数据中,确定更加有效、可靠的含水层参数,具有十分重要的意义。

受人类学习原则启发,2009 年, Bengio 等^[5]提出一种全新的机器学习框架,其核心思想是逐渐将

样本序列加入学习,简单样本先学习,较难的样本后学习。这种先易后难的学习策略对应于人类(及其他动物)在不同成长阶段学习不同难度知识的规律。2010 年, Kumar 等^[6]提出了自步学习(self-paced learning, SPL),并将其应用到潜变量模型中。之后 SPL 框架被广泛应用到各个领域^[7-10]。2015 年, Zhao 等^[11]将选择样本中的二值权值推广到实值权重,使得学习方式更加灵活。本文构造基于差分进化(differential evolution, DE)算法优化的自步学习模型,并将其应用于含水层参数确定中,可为含水层估计提供方法支持。

1 自步学习

与 Bengio 等^[5]提出的启发式策略(curriculum learning, CL)不同, Kumar 等^[6]规定了 CL 算法的基本规则,并将其作为简单的自步学习方法。一般地,

包含若干观测样本 x_i 及其标签 y_i 的训练集 $D = \{(x_i, y_i)\}_{i=1}^n$, 定义 $L(y_i, g(x_i, \theta))$ 为实测数据和理论估计数据的损失函数, 其中, θ 为函数 g 的模型参数。SPL 模型包含一个加权损失项和一个一般自步正则项:

$$\min_{\theta, v \in [0, 1]^n} E(\theta, v, k) = \sum_{i=1}^n (v_i L(y_i, g(x_i, \theta))) + f(v, \lambda) \quad (1)$$

式中 λ 为控制学习步伐的年龄系数。 $f(v, \lambda)$ 为自步正则项^[11-12], 其定义如下: 假设 v 为权值变量, l 为损失, 如果: ① $f(v, \lambda)$ 是关于 $v \in [0, 1]$ 的凸函数; ② $v^*(\lambda; l)$ 关于 l 单调递减, 并且 $\lim_{l \rightarrow 0} v^*(\lambda; l) = 1$, $\lim_{l \rightarrow \infty} v^*(\lambda; l) = 0$; ③ $v^*(\lambda; l)$ 关于 λ 单调递增, 并且 $\lim_{\lambda \rightarrow \infty} v^*(\lambda; l) \leq 1$, $\lim_{\lambda \rightarrow 0} v^*(\lambda; l) = 0$; 其中 $v^*(\lambda; l) = \arg \min_{v \in [0, 1]} vl + f(v, \lambda)$, $v \in [0, 1]$ 。定义条件②表明, SPL 模型倾向于选择简单 (损失函数值较小) 的样本。一般地, 当 λ 很小时, 模型只考虑小损失样本, 随着模型“年龄” (由年龄系数 λ 控制) 不断增大, 模型包含更多更复杂的样本, 趋于成熟。由于软权重 (实数权值) 可以更有效地反映样本在训练过程中的可信度^[11], 所以本文采用软权重和软正则项进行, 其表达式为

$$f(v, \lambda) = \frac{\gamma^2}{v + \gamma/\lambda} \quad (2)$$

式中引入参数 $\gamma > 0$ 来控制用于选取样本的权重的强度。

利用泰斯公式及其近似表达式^[4], 可计算在含水层中距抽水井距离为 r 处的水位降深 s_{ei} 。在应用自步学习时, 要求待估计的参数值能使下式表示的目标函数达到最小:

$$\min_{\theta, v \in [0, 1]^n} E(\theta, v, \lambda) = \sum_{i=1}^n v_i (s_{oi} - s_{ei})^2 + \sum_{i=1}^n \frac{\gamma^2}{v_i + \gamma/\lambda} \quad (3)$$

式中: s_{oi} 为抽水开始后第 i 个时刻观察到的实际水位降深; θ 为待估参数。

2 DE 算法优化的自步学习模型

2.1 基本思想

DE 算法是一种全局并行优化算法, 具有记忆个体最优解和种群内部信息共享的特点^[13], 这种算法主要包括交叉、变异和选择等操作算子。其基本思想是运用当前种群个体的差分变异和交叉重组得到的中间种群, 然后基于贪婪选择策略对父代种群和中间种群中的个体进行选择操作得到具有更佳适应度的子代个体。

将式 (3) 作为目标函数, 对待估参数向量 $\theta =$

(T, μ) 、样本的权重 (v_1, v_2, \dots, v_n) 进行编码得到 $(\theta, v_1, v_2, \dots, v_n)$ 作为决策变量, 利用 DE 算法进行优化。同时, 年龄系数随着迭代次数增加而不断增大。这使得在迭代前期, 算法结果不精确的情况下, 算法选择简单样本进行训练; 在迭代后期, 算法将考虑更多更为复杂的样本, 从而趋向成熟。

2.2 算法步骤及流程

步骤 1: 初始化种群规模 N 、交叉因子 C 、输入求解参数的上界和下界、最大迭代次数 k_{\max} 以及满足精度要求的迭代次数 k 。

步骤 2: 利用式 (3) 计算各个体的适应度 f_i 及其对应位置 g_i , 从其中确定全局最优适应度 f 及其对应位置 g 。

步骤 3: 当 $k < k_{\max}$ 时, 利用 $\lambda = \lambda_{\min} + \frac{(\lambda_{\max} - \lambda_{\min})k}{k_{\max}}$

更新年龄系数 λ 。

步骤 4: 由 DE 算法得到变异个体 $m_{i, k+1}$, 试验个体 $u_{i, k+1}$ 和下一代个体 $x_{i, k+1}$ 。计算并比较 $u_{i, k+1}$ 与 $x_{i, k+1}$ 相应的适应度值, 将新的最优位置赋给 $x_{i, k+1}$; 并且在新的历史最优位置中确定全局最优位置 g 。

步骤 5: 如果满足终止条件, 则停止迭代; 否则 $k = k + 1$, 转向步骤 3。

步骤 6: 输出全局最优适应度 f 及其相应位置 g 。

算法流程图如图 1 所示。

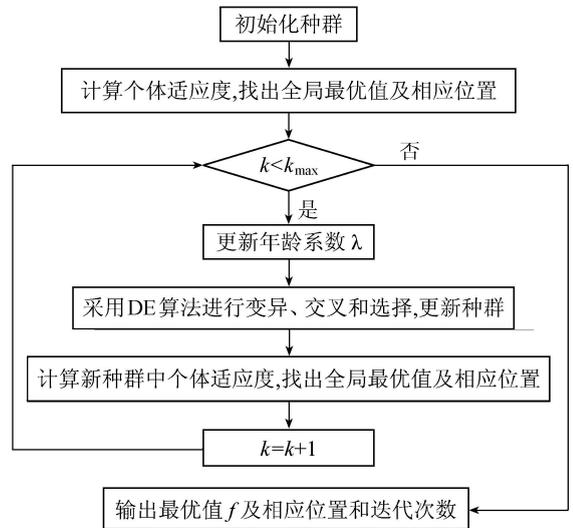


图 1 DE 算法优化的自步学习方法流程

3 数值试验

3.1 试验一

为验证本文方法的有效性, 采用文献 [2] 的数据作为原始数据进行验证试验, 试验中抽水流量为 $0.073 \text{ m}^3/\text{s}$, 持续时间为 800 min 。含水层类型、含水层介质、抽水试验的基本情况等水文地质条件详见文

献[2]。表1为距抽水井30.48 m处观察孔中的水位降深观测数据。由于测量误差不可避免,给此原始数据添加较小的测量误差。考虑到实际情况,前50%数据添加服从 $N(0,1)$ 分布的高斯误差,后50%添加服从 $N(0,10)$ 分布的高斯误差。利用MATLAB语言编程进行数值试验。将采用DE算法的最小二乘法和本文方法进行对比,两种方法均采用相同参数($N=80, k_{\max}=3000, C_R=0.8$)进行求解,运行20次,所得参数 T 和 μ 均取20次结果的平均值。所得参数值代入式(3)中,模拟水位降深结果如图2所示。引入残差平方和 R_s 对模拟结果进行评价:

$$R_s = \sum_{i=1}^n (s_{Ri} - s_{ei})^2 \quad (4)$$

式中: s_{Ri} 为抽水开始后第 i 个时刻观察到的真实水位降深; s_{ei} 为利用泰斯公式及其近似表达式^[4]计算得到的第 i 个时刻水位降深。 R_s 越小,说明模拟数据越接近原始数据。

表1 水位降深观测数据

时间 t/min	水位降深 s/mm	时间 t/min	水位降深 s/mm	时间 t/min	水位降深 s/mm
5	24	70	258	360	456
10	66	80	276	420	477
15	99	90	288	480	495
20	123	100	300	540	513
25	150	110	312	600	519
30	165	120	321	660	531
40	198	180	372	720	543
50	219	240	405	800	558
60	240	300	435		

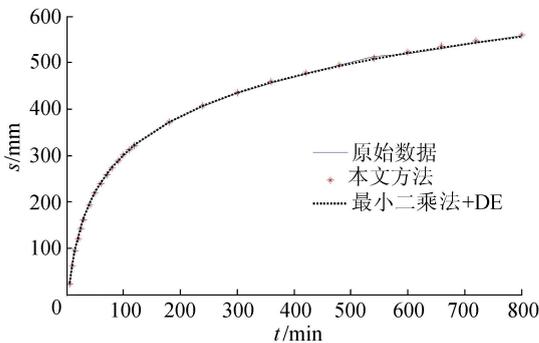


图2 试验结果对比(小误差)

在实测数据包含较小误差的情况下的试验结果对比如图2所示。由图2可知,本文方法与最小二乘法在相同的条件下,模拟数据与原始数据非常接近。将不同方法的计算结果列于表2。可以看出,两种方法所求参数值、残差平方和均相差不多。这表明,在误差较小的情况下,本文方法与最小二乘法均可以有效求解含水层参数。

3.2 试验二

考虑实际数据包含较大误差的情况,将原始数据前50%添加服从 $N(0,10)$ 分布的高斯误差,后

50%添加服从 $N(0,100)$ 分布的高斯误差,然后随机选取约40%(10个)的数据作为缺失数据。在与试验一相同试验条件下,分别以最小二乘法和本文方法进行对比试验,所得模拟水位降深结果如图3所示。

表2 不同方法计算结果

方 法	计算结果		
	$T/(\text{m}^2 \cdot \text{min}^{-1})$	μ	R_s
Sushil 方法 ^[3]	2.76	0.060	8.53×10^{-5}
直线图解法 ^[5]	3.00	0.067	5.47×10^{-4}
混沌粒子群优化算法 ^[4]	2.80	0.064	1.41×10^{-4}
本文方法(小误差)	2.82	0.062	9.47×10^{-5}
最小二乘法+DE(小误差)	2.80	0.063	1.18×10^{-4}
本文方法(大误差)	2.84	0.061	2.91×10^{-4}
最小二乘法+DE(大误差)	3.38	0.039	1.49×10^{-2}
本文方法(混合误差)	2.80	0.066	7.82×10^{-4}
最小二乘法+DE(混合误差)	2.51	0.077	3.80×10^{-3}

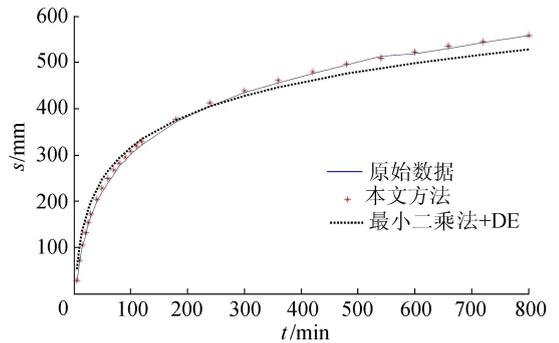


图3 试验结果对比(大误差)

由图3可以看出,当数据包较大误差时,最小二乘法计算结果严重偏离实际情况,而本文方法仍然可以较好地逼近原始数据。由表2可知,在参数值和残差平方和方面,本文方法均得到较好的结果。这表明,本文方法在数据误差较大时,可以有效地减小误差对计算结果的影响,保持较好的稳定性。

3.3 试验三

考虑到实际数据可能包含随机混合误差,将原始数据随机选取35%添加服从 $N(0,1)$ 分布的高斯误差,随机选取35%添加服从 $N(0,10)$ 分布的高斯误差,随机选取30%添加服从 $N(0,100)$ 分布的高斯误差,然后随机选取约40%(10个)的数据作为缺失数据。在与试验一相同试验条件下,分别以最小二乘法和本文方法进行对比试验,所得模拟水位降深结果如图4所示。

由图4可以看出,当测量数据包含随机混合误差时,最小二乘法方法计算结果和本文方法均有偏差,但本文方法更加逼近原始数据。表2中,本文方法所求参数值和残差平方和均接近传统方法并且优于最小二乘法。这表明,本文方法在实测数据包含混合随机误差时,可以有效地减少误差对计算结果的影响,保持较好的稳定性。

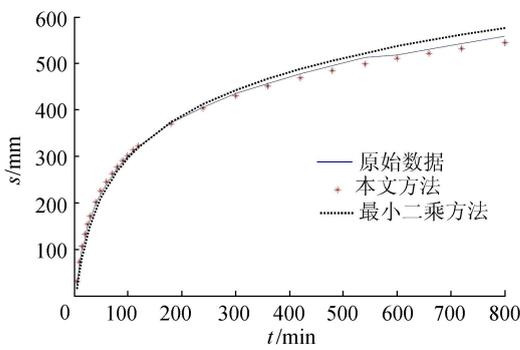


图4 试验结果对比(混合误差)

4 结论

针对以往确定含水层参数的研究中没有考虑误差影响的情况,本文尝试将机器学习中自步学习方法引入到确定含水层参数中,构造了基于差分进化算法优化的自步学习方法,数值试验结果表明:①在误差较小的情况下,自步学习方法的精度与最小二乘法相当的结果;②在误差较大的情况下,自步学习方法可以有效减少误差的影响,使得所求参数更加接近真实情况;③在实测数据包含随机混合误差情况下,此方法同样可以达到满意结果。因此,自步学习方法可以有效确定含水层参数,并保持良好稳定性。

参考文献:

[1] SUSHIL K S. Confined aquifer parameters from temporal derivative of drawdowns [J]. Journal of Hydraulic Engineering, 2001, 127 (6): 466-470.

[2] U. S. Department of the Interior (USDI). Groundwater Manual [M]. Washington, D. C. : U. S. Govment Printing Office, 1977.

[3] 江思珉,蔡奕,王敏,等. 基于和声搜索算法的地下水污染源与未知含水层参数的同步反演研究[J]. 水利学报, 2013, 43 (12): 1470-1477. (JIANG Simin, CAI Yi, WANG Min, et al. Simultaneous identification of groundwater contaminant source and aquifer parameters by harmony search algorithm [J]. Journal of Hydraulic Engineering, 2013, 43(12): 1470-1477. (in Chinese)).

[4] 周秀秀,常安定,郭建青,等. 混沌粒子群优化算法在确定含水层参数中的应用[J]. 水资源与水工程学报, 2013, 24 (1): 96-99. (ZHOU Xiuxiu, CHANG Anding, GUO Jianqing, et al. Application of chaos particle swarm optimization algorithms to estimation of aquifer parameters [J]. Journal of Water Resources & Water Engineering, 2013, 24(1): 96-99. (in Chinese))

[5] BENGIO Y, LOURADOUR J, COLLOBERT R, et al. Curriculum learning [C]//International Conference on Machine Learning. New York: ACM, 2009: 41-48.

[6] KUMAR M P, PACKER B, KOLLER D. Self-paced

learning for latent variable models [C]//Proc of the Neural Information Processing Systems, New York: Curran Associates, 2010: 1189-1197.

[7] TANG K, RAMANATHAN V, LI F F, et al. Shifting weights: adapting object detectors from image to video [J]. Advances in Neural Information Processing Systems, 2012: 638-646.

[8] KUMAR M P, TURKI H, DAN P, et al. Learning specific-class segmentation from diverse data [J]. Proceedings, 2011, 58 (11): 1800-1807.

[9] YONG J L, GRAUMAN K. Learning the easy things first: self-paced visual category discovery [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2011: 1721-1728.

[10] Supancic Iii J S, RAMANAN D. Self-paced learning for long-term tracking [J] Computer Vision & Pattern Recognition, 2013, 9(4): 2379-2386

[11] ZHAO Q, MENG D, JIANG L, et al. Self-paced learning for matrix factorization [C]//Twenty-Ninth AAAI Conference on Artificial Intelligence. California: AAAI Press, 2015.

[12] JIANG L, MENG D, MITAMURA T, et al. Easy samples first: self-paced reranking for zero-example multimedia search [C]//ACM International Conference on Multimedia. New York: ACM, 2014: 547-556.

[13] 王晓晨,常安定,王静云,等. 各向异性含水层参数估计的差分进化算法[J]. 西北农林科技大学学报(自然科学版), 2015, 43 (4): 223-228. (WANG Xiaochen, CHANG Anding, WANG Jingyun, et al. Differential evolution method for estimating the anisotropy aquifer parameters [J]. Journal of Northwest A&F University (Natural Science Edition), 2015, 43 (4): 223-228. (in Chinese))

(收稿日期: 2016-08-12 编辑: 郑孝宇)

(上接第 50 页)

[12] 李成业,练继建,刘昉,等. EMD 与小波阈值联合滤波方法的改进及其在泄流结构振动分析中的应用[J]. 振动与冲击, 2013 (19): 63-70. (LI Chengye, LIAN Jijian, LIU Fang, et al. An improved filtering method based on EMD and wavelet-threshold and its application in vibration analysis for a flood discharge structure [J]. Journal of Vibration and Shock, 2013 (19): 63-70. (in Chinese))

[13] 徐国宾,瞿娟,高仕赵. 平面闸门启闭过程中的动水垂直力数值模拟研究[J]. 水电能源科学, 2012 (10): 132-135. (XU Guobin, ZI Juan, GAO Shizhao. Numerical simulation of hydrodynamics vertical force during plane gate opening and closing process [J]. Water Resources and Power, 2012 (10): 132-135. (in Chinese))

(收稿日期: 2016-09-06 编辑: 张志琴)