基于 MCS 的 SVM 参数优化研究

郭一格†，孙　力，刘以安
（江南大学 物联网工程学院，江苏 无锡 214122）

摘　要：由于优化杜鹃算法是利用了鸟类特殊的利维飞行模式的群体智能算法，并且增加了粒子间的信息交流，故将该算法引入支持向量机惩罚系数和核参数的自动寻优中。给出了实现方式，并讨论了概率参数的设置对收敛性的影响。通过与传统的 GA/PSO-SVM 对比验证，MCS-SVM 方法使得分类精确率平均提高 2.28%，既能显著提高分类效率，又表现出很好的泛化性能。

关键词：支持向量机；参数寻优；同步智能寻优；优化杜鹃搜索
中图分类号：TP301.6　　**文献标志码**：A　　**文章编号**：1001-3695(2012)12-4553-03
doi:10.3969/j.issn.1001-3695.2012.12.039

Parameters optimization research of SVM based on MCS

GUO Yi-ge†, SUN Li, LIU Yi-an
(*School of Internet of Things Engineering*, *Jiangnan University*, *Wuxi Jiangsu* 214122, *China*)

Abstract: The MCS is a swarm intelligent alogrithm which is based on the Lévy flight behavior of some birds. It also increases the exchange of information between particles. So it could be introduced into the parameters optimization problem of penalty parameter C and Kernal parameter σ of SVM. This paper gave the implementation and discussed the influence of probability parameters on convergence. The simulation results show it can get 2.28% improvement in average with this method, then it demonstrates that thiss method outperforms the methods such as GA/PSO-SVM by not only significantly giving more accurate results but also improving generalization ability.

Key words: support vector machine(SVM); parameters optimization; synchronous intelligent optimization; modified cuckoo search(MCS)

0　引言

支持向量机(SVM)模型的两个重要参数(即惩罚系数 C 和核参数)选取的好坏直接影响应用上的分类精度。任意给定或凭测试经验给定都不足为据。目前大多运用智能优化算法对其进行同步优化操作，并对训练样本进行迭代训练自动选择，如 Leung 等人[1] 提出了基于实值遗传算法(genetic algorithm，GA)实现了 SVM 模型参数的自动选择，该方法基于遗传算法的隐含并行高效性和全局最优的性能选择了 SVM 模型参数，提高了 SVM 的构造效率，而且进一步提高了分类器的识别率。吴景龙等人[2,3]也采用了遗传算法优化，最终能够有效解决此问题，但收敛速度较慢。而与 GA 相比，粒子群优化(PSO)算法具有执行速度快、受问题维数变化影响小等优点，且容易实现。因此，Huang 等人[4~6]提出将 PSO 算法应用到支持向量机参数的选择中，但是这些算法存在适应度函数的局部适应性，或者没有考虑到生物之间通信的可能性等局限性。而本文研究一种名为 MCS[7] 的群体智能生物学习方法，该方法结合了许多鸟类及果蝇特殊的利维飞行模式进行搜索，同时增加了群体之间的信息交流，加速收敛效率。

而对于任何搜索算法，设置合理的搜索范围才能保证快速高效地找到最优参数组合，因此本文从 SVM 分类中核参数对于支持向量覆盖效能影响出发，对核参数取值范围进行合理预估计，而惩罚系数暂且按照经验设置取值范围。然后在这个参数空间引入 MCS 方法实现 SVM 参数寻优，将它称为 MCS-SVM 方法，并且对运用到 SVM 的参数寻优所需要的参数设置也进行了考量。通过仿真与传统 GA-SVM/PSO-SVM 方法作对比，验证了所提出方法的有效性。考虑 RBF 核函数的广泛应用，本文仅针对 RBF 核进行研究。

1　SVM 同步优化

1.1　SVM 基本思想

假设存在一组带有类别标记的训练样本 $T=\{(x_1,y_1),\cdots,(x_n,y_n)\}$，$x_i$ 是 n 维向量，y_i 代表 x_i 所属类。其中每个输入点 $x_i\in R^d(i=1,\cdots,n)$ 属于两类中的某一类，其类别标记为 $y\in\{+1,-1\}$。SVM 的主要目的是构造一分类超平面(最优超平面)以分割两类不同的样本，使得分类间隔(margin)最大。最终归结为约束条件下求解二次规划问题：

$$\min \frac{1}{2}(w\times w)+C\sum_{i=1}^{n}\varepsilon_i$$
$$\text{s.t.}\quad y_i[(w\times x_i)+b]\geqslant 1-\varepsilon_i$$
$$\varepsilon_i\geqslant 0,i=1,\cdots,n \tag{1}$$

其中：C 为惩罚因子，ε 是线性不可分时引入的松弛因子。

收稿日期：2012-05-02；**修回日期**：2012-06-18

作者简介：郭一格(1988-)，女(通信作者)，山东青岛人，硕士研究生，主要研究方向为信息融合与雷达对抗(guoyi6ge@163.com)；孙力(1966-)，男，教授，硕导，博士，主要研究方向为多媒体技术、数据搜索技术；刘以安(1963-)，男，教授，硕导，博士，主要研究方向为数据融合、雷达对抗、模式识别与智能系统.

引入 Lagrange 乘子 $\alpha_i(i=1,\cdots,n)$ 得其对偶形式：

$$\max \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j$$

$$\text{s.t.} \quad 0 \leqslant \alpha_i \leqslant C, i=1,\cdots,n$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0 \tag{2}$$

在非线性的情况下，引入核函数映射：

$$K(x_i, y_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$$

则式(2)变为

$$\max \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, y_j)$$

由以上推导得到以下决策函数：

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i y_i K(x, x_i) + b\right) \tag{3}$$

其中；$K(x,y)$ 为核函数；b 是根据训练样本确定的阈值；α_i 由二次规划确定；n 为训练样本数。

对 RBF 核有

$$K(x_i, y_j) = e^{-\|x_i - x_j\|^2/2\sigma^2} \tag{4}$$

1.2 优化杜鹃搜索

2009 年 Yang 等人[8] 提出了一种元启发式搜索算法(cuckoo search，CS)，该算法基于杜鹃种群寄生繁衍策略，并结合了许多鸟类及果蝇特殊的利维飞行模式，而 MCS 是基于 CS 算法的改进。Walton 等人[7] 已经通过对多种基准函数进行对比实验，验证了使用 MCS 比使用传统智能优化算法 GA/PSO 等更加有效。

CS 基于以下三种规则：a)每个杜鹃每次下一个蛋，代表一种并列的解决方案，随机放在一个鸟巢中；b)一部分鸟巢放着优质蛋，即好的解决方案，这些鸟巢将会传到下一代；c)鸟巢的数量是固定的，杜鹃蛋被寄主鸟发现的概率是 $Pa \in (0,1)$。一旦某个鸟巢被发现，寄主鸟就丢弃鸟蛋或鸟巢，以免影响寻找最佳解决方案，同时产生新的鸟巢。

如果进行足够的计算，CS 总能够找到优化方案[9]，但因为搜索完全依赖随机过程，就不能够保证快速收敛。MCS 改进了利维分布(Lévy flight)[8,9] 的步长 α，这样一来，粒子或者蛋就会在逼近最佳解决方案过程中进行局部搜索。另外，增加鸟蛋间的信息交流，提高了收敛率，从而使得这个搜索方法应用范围更广。

MCS 另一优势是只需要调节参数，即鸟巢被抛弃概率(PA)和优质鸟巢的修复概率($1-PA$)。

2 基于 MCS 的 SVM 参数同步优化

2.1 参数区间预估计

对于采用 RBF 核函数的 SVM，惩罚系数 $C>0$ 和核参数 $\sigma>0$。这两个参数的重要作用为：惩罚系数 C 表示对错误分类的惩罚指数，值太大会导致分类准确率在训练过程中很高而在测试过程中很低，反之则分类准确率不可靠，模型无效；核参数 σ 对分类结果更重要，因为它的值影响在特征空间的分割结果，值过大会导致过拟合，反之不成比例的过小值会导致欠拟合。

由于核函数也可以理解为样本对其附近样本的覆盖，所以核参数 σ 可以参考样本间的距离。对所有的训练样本，限制 σ 的范围：

$$k_1 d_{\min} = \sigma_l < \sigma < \sigma_u = k_2 d_{\max} \tag{5}$$

其中：$d_{\min} = \frac{1}{l}\sum_{i=1}^{l} d_i^{\min}$ 和 $d_{\max} = \frac{1}{l}\sum_{i=1}^{l} d_i^{\max}$ 分别表示训练样本的平均最近距离和平均最远距离；$d_i^{\min} = \min_{i,j\neq i} d_{ij}$ 和 $d_i^{\max} = \max_{i,j\neq i} d_{ij}$ 表示每个样本的最近样本距和最远样本距，这里采用欧氏距离：$d_{ij} = d(x_i, x_j) = \| x_i - x_j \|$。

对于 k_1 与 k_2 的设定讨论如下：

a)下界 σ_l。认为小于 0.003(小概率事件)的支持向量难以覆盖其相邻样本，而小于这个数值的核函数数值说明当前样本对其相邻样本影响很微小，即核函数的半径不宜小于此值，否则容易造成过适应。令 $K(x_i, y_j) = 0.003$，根据式(4)可得 $\sigma_l = 0.2934 d_{\min}$。

b)上界 σ_u。支持向量应该对距离其近的向量影响大，而对距离较远的向量影响小。较大的 σ 覆盖范围大，但是至少该核函数对距离最远的向量能够体现出这种影响的衰减。设定衰减系数 r，则有 $e^{-d_{\min}^2/2\sigma^2} - e^{-d_{\max}^2/2\sigma^2} < r$。令 $r = 0.003$，并假设 $d_{\min} << d_{\max}$，则 $\sigma_u = 12.9 d_{\max}$。

据以上分析，本文近似取 $k_1 = 0.3$，$k_2 = 13$，带入式(5)，即可确定 $\sigma \in (\sigma_l, \sigma_u)$。本文也通过实验采用 MCS-SVM 分类方法测试，对几种数据库实验均能得到如图 1 所示的效果，因而得出结论，这种核参数的设置比更大的范围(如[0.001, 1000])，收敛速度大大提高。参数 C 的范围不易确定，在多次测试过程中观察到设置太大通常效果也不好，按经验设为 $C \in [C_l, C_u] = [0.1, 100]$。

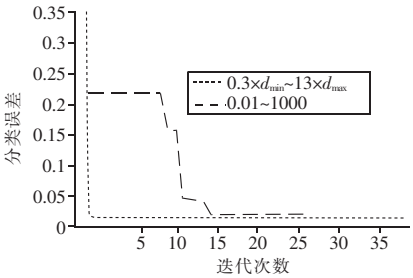

图 1 核参数预估计范围对比

2.2 MCS-SVM 分类步骤

总结 MCS-SVM 混合方法实现步骤如下：

a)对训练与测试样本统一归一化，用式(5)预估计 σ 取值范围，并设置 C 取值范围，设利维飞行的最大步数 $A=1$。

b)以拉丁抽立方方法初始化 20 个二维的鸟巢 $D = \{x_i\} = \{[C_i, \sigma_i]\}$，$i \in (1,2,\cdots,20)$，计每组 x_i 下对训练样本的验证分类准确率 F_i 作为寻优的目标函数适应值；初始化 MCS 的概率参数：$PA = 0.75$。

c)从 $G=1$ 起迭代 200 次步骤 d)：将 x_i 按 F_i 从大到小排序：

(a)选出需要抛弃的鸟巢 D_1，当前鸟巢为 $x_i(i=1,2,\cdots,(20\times PA))$；计算利维飞行步大小：$\alpha \leftarrow A/\sqrt{G}$；由利维飞行得新的鸟巢 x_k 用予代替 $x_i \leftarrow x_k$，并计算它们的适应值 F_i；

(b)剩下的鸟巢则为适应值高的鸟巢，作为优质鸟巢 D_2，当前鸟巢为 $x_i(i=1,2,\cdots,20\times(1-PA))$；从 D_2 中随机选择 x_j；

如果 $x_i = x_j$，

计算利维飞行步大小：$\alpha \leftarrow A/G^2$；

由利维飞行得新鸟巢 x_k,并计适应值 F_k;

从 D_2 中随机选择鸟巢 x_l,并计适应值 F_l;

如果 $F_k > F_l, x_l = x_k, F_l = F_k$;

反之如果 $x_i \neq x_j$,

按照黄金比例,求得新鸟巢位置距离:$dx = |x_i - x_j| / \varphi$;//黄金比率 $\varphi = 0.5 \times (1 + \sqrt{5})$;

x_i 朝向 x_j 移动 dx 距离,得到新的鸟巢 x_k,并计适应值 F_k;

从 D_2 中随机选择鸟巢 x_l,并计算适应值 F_l;

如果 $F_k > F_l, x_l = x_k, F_l = F_k$;

d)最终取使得训练样本分类准确率最高的那组 C 和 σ 作为最佳参数,在测试样本中测试。

3　实验及结果

本文采用 UCI 数据库中的四个数据库验证,如表 1 所示。为既能保证实验效率又保证准确率,实验所采用的 GA、PSO、MCS 寻优操作中种群数或者鸟巢数量均为 20,迭代次数为 200。C 和 σ 的取值范围同上文。

表 1　数据库信息

标记	数据库	训练样本	测试样本	属性维数
1	learnsample	440	129	3 维
2	heart_disease	500	190	14 维
3	wine_divide	119	59	13 维
4	waves	80	187	44 维

3.1　MCS 运用到 SVM 概率参数分析

MCS 只需要调节鸟巢被抛弃概率(PA)和优质鸟巢的修复概率($1 - PA$)两个参数,本节对 SVM 参数优化过程中 MCS 的概率参数设置进行了实验验证。

将概率参数 PA 在[0.01,1]内依次线性选取 10 个值进行对比,选取 wine_divide 数据库和 heart_disease 数据库,分别运行三次实验。选取 6 个特征明显的概率参数讨论实验结果所得各项效果值,因实验发现不同的参数设置导致收敛稳定性不同,稳定为正,如表 2 所示。

表 2　MCS 概率参数对 SVM 分类效果比较

PA	稳定比例	训练误差	测试误差	C 值	σ 值
0.12	5/7	0.4200	0.2778	57.5425	25.0098
0.34	8/4	0.2000	0.2111	18.8949	9.9738
0.56	8/4	0.1400	0.1778	94.0175	5.7037
0.65	9/3	0.1400	0.1778	55.8323	5.3465
0.78	11/1	0.1000	0.1333	56.2780	0.0203
0.89	9/3	0.1133	0.1556	40.3660	0.0186

从表 2 的结果中可见,$PA = 0.78$ 时,12 次实验中有 11 次是稳定收敛,稳定性明显最能保证,而所得训练误差及测试误差都明显小。

迭代训练过程中分类误差收敛速率分别可在图 2 和 3 中观察到,可以看出相比之下参数设置为 0.67 与 0.89 时的收敛速度都容易保证,但稳定性还是不如 0.78。而文献[10]中对许多基准问题进行了测试,得出结论 PA 设为 0.75 时对大部分函数搜索效果最佳。最终,本实验也采取了这种 MCS 的参数设置方案。

3.2　MCS-SVM 实际分类效果实验分析

利用 MCS-SVM 方法对四种数据库进行验证分析,所选用的 4 个数据库对训练数据和测试数据进行分类预测所得到的分类准确率,实验结果如表 3 所示。

表 3　分别用 GA/PSO/MCS 所得分类效率　　/%

NO		GA-SVM	PSO-SVM	MCS-SVM	突出
1	train	99.0909	99.0909	99.0909	–
	test	75.969	75.969	76.7442	0.78
2	train	85.4	85.4	88.8	3.40
	test	85.7895	85.7895	86.3158	0.53
3	train	43.1034	98.3193	100	1.68
	test	43.1034	98.3051	100	1.69
4	train	83.75	83.75	83.75	–
	test	66.8449	66.8449	77.0053	10.16
					2.28

从表 3 后一列即可看出:无论训练迭代寻优的最终分类精确度(train 行),还是用训练得到的 C 和 σ 对测试样本进行测试所得分类精确度(test 行),MCS-SVM 几乎都优于 PSO-SVM,统计得本文实验所得分类准确率平均提高 2.28%。

而训练过程中对训练样本会进行迭代寻优操作,从迭代过程中 SVM 适应值可以看出 MCS 能够保证收敛,能够找到对训练样本分类准确率更高的 SVM 参数组合,如图 4、5 所示。

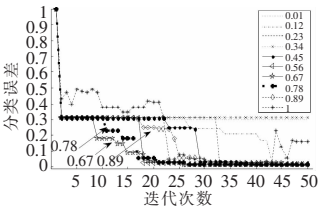

图 2　wine 数据库,概率参数取值对比结果

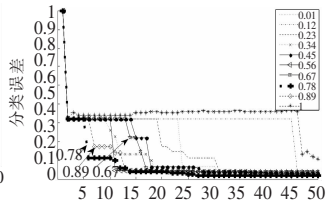

图 3　heart 数据库,概率参数取值对比结果

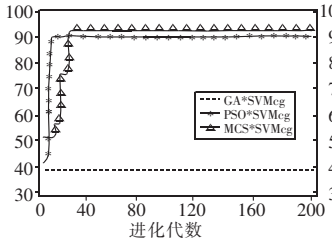

图 4　对 wine 数据库训练,迭代寻优过程

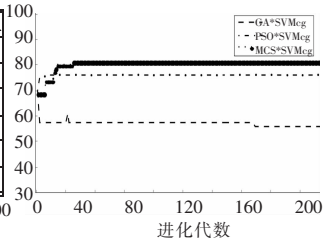

图 5　对 waves 数据库训练,迭代寻优过程

实验中观察发现,随着问题维数的增加,MCS 优于 PSO 的差距就越大,如图 4 与 5 中的对比,说明 MCS 对于复杂数据库的处理能力也优于 PSO 等方法。

4　结束语

本研究提出一种基于优化杜鹃搜索的支持向量机新型自动分类模型,其优越性在于既能够细化局部搜索,又能够保证全局搜索。这也是因为 MCS 增加优质粒子之间的信息交流,全局搜索体现在它能够及时放弃劣质粒子,并且以利维飞行模式搜索。MCS 这些优势的实现无须更大的时间代价,很容易实现,并且需要调节的参数极少。仿真结果验证了所提出的方法能提高模型的分类准确率。另外,根据 RBF 核函数特点,通过样本间距得出核参数取值范围,不但减小了模型优化算法的搜索空间,同时提高了算法寻优能力和加快收敛。

参考文献:

[1] LEUNG Y W, WANG Yu-ping. An orthogonal genetic algorithm with quantization for global numerical optimization[J]. IEEE Trans on Evolutionary Computation, 2001, 5(1): 41-53.

(下转第 4559 页)

表 2 中特征按重要性由大到小排序,可以看出两种算法得到的特征排序是有差别的,两种算法所消耗的时间相差很大,基于核空间类间平均距的特征选择算法比 SVM-RFE 算法执行得更快。

按照表 2 的特征排序,依次取前若干个特征组成不同维数的样本集,5 折交叉验证正确率与样本集维数的关系如图 4 和 5 所示。

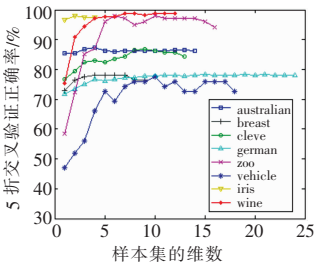

图 4　SVM-RFE 算法

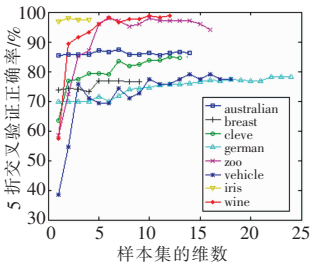

图 5　基于核空间类间平均距算法

最优特征子集就是图 4 和 5 中 5 折交叉验证正确率最大时的样本集对应的特征子集,两种算法得到的最优特征子集的特征个数及 5 折交叉验证识别率如表 3 所示。

表 3　特征选择结果

数据集	特征数	识别率/%	SVM-RFE		核空间类间平均距	
			特征数	识别率/%	特征数	识别率/%
australian	14	86.38	4	87.39	7	87.25
breast	9	76.53	4	77.98	5	76.90
cleve	13	84.46	9	86.82	11	84.80
german	24	78.10	15	78.30	16	78.24
zoo	16	94.06	6	98.02	6	98.02
vehicle	18	77.42	10	77.42	14	79.03
iris	4	97.33	2	98.00	2	98.00
wine	12	98.88	7	98.88	6	98.88

表 3 数据表明,基于核空间类间平均距的特征选择算法是正确的、有效的,该算法可以从样本集中选择出对分类识别有用的特征,去除冗余特征和干扰特征,使识别率得到提高;基于核空间类间平均距的特征选择算法得到的最优特征子集与 SVM-RFE 算法得到的最优特征子集是有差别的,而且最优特征子集的特征个数也稍多,但是基于核空间类间平均距的特征选择算法比 SVM-RFE 算法运算效率要高很多,因此本文的算法更具实用性。

5　结束语

本文分析了RBF核参数对样本数据在核空间分布的影响,给出了核参数与样本核空间类间平均距的关系,提出了基于核空间类间平均距的特征选择算法,以样本中各个特征对核空间类间平均距贡献的大小为据进行特征重要性排序,在 UCI 数据集进行的实验验证了该算法的正确性、有效性,而且该算法比 SVM-RFE 算法进行特征选择所消耗的时间更短。

参考文献:

[1] SAEYS Y, INZA I, LARRANAGA P. A review of feature selection techniques in bioinformatics [J]. Bioinformatics, 2007, 23 (19): 2507-2517.

[2] KIRA K, RENDELL L. The feature selection problem: traditional methods and a new algorithm A[C]//Proc of the 9th National Conference on Artificial Intelligence C. New Orleans: AAAI Press, 1992: 129-134.

[3] KONONENKO I. Estimating attributes: analysis and extensions of RELIEF [C]//Proc of European Conference on Machine Learning. 1994: 171-182.

[4] GUYON I, WESTON J, BARNHILL S, *et al*. Gene selection for cancer classification using support vector machines [J]. Machine Learning, 2002, 46(13): 389-422.

[5] DING Yuan-yuan, WILKINS D. Improving the performance of SVM-RFE to select genes in microarray data[J]. BMC Bioinformatics, 2006, 7(S2): 12.

[6] MAO Yong, PI Dao-ying, LIU Yu-ming, *et al*. Accelerated recursive feature elimination based on support vector machine for key variable identification[J]. Chinese Journal of Chemical Engineering, 2006, 14(1): 65-72.

[7] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.

[8] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42.

[9] HSU C W, CHANG C C, LIN C J. A practical guide to support vector classification[EB/OL]. (2003). http://www.cise.ntu.edu.tw/cjlin/papers/guide/guide.pdf.

[10] 宋小杉, 蒋晓瑜, 罗建华, 等. 基于类间距的径向基函数支持向量机核参数评价方法分析[J]. 兵工学报, 2012, 33(2): 203-208.

[11] BLAKE C L, MERZ C J. UCI Repository of machine learning databases [EB/OL]. (2003). http://www.ics.uci.edu/~mlearn/MLRepository.html.

(上接第 4555 页)

[2] 吴景龙, 杨淑霞, 刘承水. 基于遗传算法优化参数的支持向量机短期负荷预测方法[J]. 中南大学学报: 自然科学版, 2009, 40(1): 180-184.

[3] OSOWSKI S, SIROIC R, MARKIEWICZ T, *et al*. Application of support vector machine and genetic algorithm for improved blood cell recognition[J]. Instrumentation and Measurement, 2009, 58(7): 2159-2168.

[4] HUANG C L, DUN J F. A distributed PSO-SVM hybrid system with feature selection and parameter optimization[J]. Applied Soft Computing, 2008, 8(4): 1381-1391.

[5] 任江涛, 赵少东, 许盛灿, 等. 基于二进制 PSO 算法的特征选择及 SVM 参数同步优化[J]. 计算机科学, 2007, 34(6): 179-182.

[6] 乔立岩, 彭喜元, 彭宇. 基于微粒群算法和支持向量机的特征子集选择方法[J]. 电子学报, 2006, 34(3): 496-498.

[7] WALTON S, HASSAN O, MORGAN K, *et al*. Modified cuckoo search: a new gradient free optimisation algorithm[J]. Chaos, Solitons & Fractals, 2011, 44(9): 710-718.

[8] YANG X S, DEB S. Cuckoo search via Lévy flights[C]//Proc of World Congress on Nature & Biologically Inspired Computing, NaBIC. 2009: 210-214.

[9] PAVLYUKEVICH I. Cooling down Lévy flights[J]. Journal of Physics A: Mathematical and Theoretical, 2007, 40: 225-232.

[10] YANG X S, DEB S. Engineering optimisation by cuckoo search[J]. International Journal of Mathematical Modelling and Numerical Optimisation, 2010, 1(4): 330-343.

[11] TUBA M, SUBOTIC M, STANAREVIC N. Modified cuckoo search algorithm for unconstrained optimization problems[C]//Proc of the 5th European Conference on European Computing Conference. 2011: 263-268.